FIDELITY ASSESSMENT FOR MODEL SELECTION (FAMS): A FRAMEWORK FOR INITIAL COMPARISON OF MULTIFIDELITY MODELING OPTIONS

A Dissertation
Presented to
The Academic Faculty

By

Adam W. Cox

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Aerospace Engineering

Georgia Institute of Technology

May 2019

FIDELITY ASSESSMENT FOR MODEL SELECTION (FAMS): A FRAMEWORK FOR INITIAL COMPARISON OF MULTIFIDELITY MODELING OPTIONS

Approved by:

Prof. Dimitri N. Mavris
School of Aerospace Engineering
Georgia Institute of Technology

Prof. Daniel P. Schrage
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Alicia M. Sudol
School of Aerospace Engineering
Georgia Institute of Technology

Dr. Stephen J. Edwards
Marshall Space Flight Center - Advanced Concepts Office
National Aeronautics and Space Administration

Mr. Philip A. Fahringer
Lockheed Martin Aeronautics Sustainment
Lockheed Martin Corporation

Date Approved: March 5, 2019

Research is what I'm doing when I don't know what I'm doing

Wernher von Braun

To my wife, Mallory Rose

Thank you taking this journey with me

ACKNOWLEDGEMENTS

I could not have gotten to this point in my schooling without the help of many individuals throughout my life. I would like to start by thanking my committee: Dr. Dimitri Mavris, Dr. Daniel Schrage, Dr. Alicia Sudol, Dr. Stephen Edwards, and Mr. Philip Fahringer. The discussions and feedback you provided helped make this work what it is, and your encouragement helped keep the ball rolling to get to this point. I particularly would like to thank Dr. Mavris, my advisor, for allowing me the opportunity to be a part of the Aerospace Systems Design Laboratory (ASDL), where I have been involved in a great breadth of research, which has helped me grow, and inspired this work. The focus on becoming a well-rounded engineer, with an emphasis on how research is applied outside academia, aligns with my own philosophy and drives me towards continual improvement.

I would also like to thank all of the research engineers that have helped me along the way, for providing the example for how to tackle difficult problems, grow, and learn. Thank you to the fellow students who have helped throughout the graduate school process, specifically my long-time officemate and work partner, Coleby Friedland, for always providing a sounding board and a push to always be better.

It would be remiss of me to ignore the impact of all of the former teachers that have inspired me, from elementary through undergraduate. The opportunities afforded me and the role models that you provided, both in and out of the classroom, have contributed a great deal to the person I am today.

I must thank my wife, Mallory, for always providing the appropriate balance of love, support, and motivation that I need to make it through. Lastly, I have to thank my family for always supporting me. I know one of you is likely to be telling someone about me at any given moment. Thank you for believing in me, driving me, and instilling the confidence to know that I can tackle whatever challenge I set for myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Model development and selection are crucial to the process of design and analysis. Ideally, model selection would entail a rigorous quantitative approach, through comparison of model data to truth data. However, if sufficient data were available to guarantee model credibility and applicability, modeling would not be needed. As such, given a problem definition, the enumeration of, and selection from, relevant modeling options relies on expert opinion. These processes are typically performed ad hoc, relying as much on familiarity and availability as on model fidelity, and the modeling options and justifications for decision-making are rarely captured.

Additionally, even if a model could be proven to be complete and perfect representation of the physical system, such a model would likely require an infeasible amount of time to run. As such, compromises in fidelity must always be made in the interest of meeting cost, or runtime, requirements. To address this, a framework is developed to provide a method for capturing expert knowledge in initial comparison of multifidelity modeling options and providing justification for decision-making in terms of both fidelity and efficiency.

Fidelity is a term that many have worked to define in a more usable manner. In the literature, resolution and abstraction have been used to describe fundamental aspects of a model that drive much of its behavior. In addition to those two attributes, scope, or how much of the system the model represents, is presented in this work as the third fundamental characteristic of fidelity. Through the comparison of these characteristics, an understanding of the relative fidelity of models can be estimated, even before model data is available. This understanding is represented by providing scores with respect to resolution, abstraction, and scope, and combining them using Kernel Density Estimation (KDE). The density estimates are used to understanding the relative comparison between the models.

As model data becomes available, it should be used to update the magnitudes of the relative fidelity assessments based on model agreement, and help to identify deficiencies

that were not previously considered, or were overlooked in verification. Based on the hypothesis that, especially for larger model sets, model agreement implies higher fidelity, correlation and error metrics are used to generate additional scores. These scores are used to update the prior density estimates using KDE.

Whether or not model data is available, the understanding of fidelity should be combined with information regarding the efficiency of models to find the non-dominated set of multifidelity combinations and compare them to the fidelity and efficiency of individual models. This can be used to justify single or multi-model selection based on the current set of fidelity and cost requirements, and should be revisited as more data is generated or requirements change. A set of notional models is used to develop the initial methods for when model data is not available. A set of I-beam finite element models (FEM) are used to verify the methods and test how correlation and error metrics can represent model agreement to adjust the fidelity estimates. From there, a full decision-making framework is developed using these methods and applied to the problem of estimating primary wing structural weight using FEM as wing aspect ratio is varied.

Using the three model sets, it is shown how fidelity should be described, relative model fidelity can be quantified, and a more informed decision-making process can occur in the early phases of a project. This also informs how models should be developed: any model responses and information about the cost to generate, analyze, or post-process, can be used in the manner described herein to justify continued model development. Additionally, the modeling options and relative fidelity descriptions can be recorded, both to inform an initial decision-making process and to allow for reconsideration if new models are considered, data is generated, or requirements change.

CHAPTER 1
INTRODUCTION AND BACKGROUND

As statistician George Box famously wrote in 1978, "All models are wrong but some are useful[1]." Similarly, it could be said that building a model is difficult, but developing and selecting the *correct* model is nearly impossible in a practical context.

1.1 Motivation

In the process of design and analysis of physical systems, models must be used to represent the system as it does not yet exist and may require a large amount of time and money to create. Designing aerospace vehicles are particularly difficult represent, even in models, due to their complexity. This complexity comes from the difficult tasks they are required to perform: carrying payloads into extreme environments via some combination of list and propulsion. As such, every bit of mass, from the primary structures to the wiring, has to buy its way onto the final product.

Due to some combination of altitude and speed, these vehicles must endure a wide range of temperatures, pressures, and forces as they centure from sea level to, in some cases, the vacuum of space. Due to these eextremes, there is not only a continual push to incorporate new technologies that improve performance, but potentially revolutionary conceptual changes that nullify the applicability of previously existing models.

The technological cutting edge and multidisciplinary nature of these systems makes it that much more difficult to instill the level of confidence that is needed to make design decisions. Any model that is used to aid the designer must be understood in great depth, especially when there is little data in place to prove the model's credibility. This problem is even more glaring in early design.

In early conceptual design, models need to be predominantly fast. Evaluation of high

level concepts, as implied by the name conceptual design, require thorough exploration of the widest variable space. For example, at this point, the decision-maker may still be weighing whether different wing-fuselage-tail configurations are appropriate, or even if a cruise missile would complete the given task better than an aircraft. Models used at this point rely less on physics or detailed logic, and are often primarily regressions of historical data. These can be fast and accurate, but are limited to the assumptions of the training data. This means that they can only interpolate, and, more than likely, if the model is used over a long span of time, the assumptions and limitations of the underlying data will be lost.

As such, early conceptual models are used to make enough decisions to guide further model development. To enable further decisions and move from conceptual to detailed design, less efficient but higher fidelity models are developed and used to generate data. This data, as well as the data from experiments, prototypes, etc., is used to make more and more specific decisions. In addition, the higher fidelity and experimental data is used to reduce uncertainty and validate previous decisions. Over time, evaluating similar concepts, a series of models can be developed through experience, moving from an emphasis on efficiency to an emphasis on fidelity as the process moves from early conceptual to late detailed design.

However, as even non-revolutionary technologies are applied, such as a new material or manufacturing process, the changes to the system can often be too much for the tools traditionally used in conceptual and even preliminary design. The aforementioned extreme conditions and cutting edge technology infusion exacerbate this issue, especially in aerospace applications. The physical and monetary scale of projects related to aerospace vehicles mean that if a problem arises during testing late into the process, the project or even the company could be in jeopardy.

Even when poor decisions are not an imminent threat to a project or company, there will always be rework. Whether the design space was not thoroughly explored or an important attribute was not adequately represented by the model, a decision was made based

on a misrepresentation of the system, which requires additional effort to correct. In fact, according to JPL data, growth in the mass and power requirements of spacecraft has typically ranged from 25-40% due to a variety of causes: design changes to the mission or the system, complexity, inheritance, technology infusion, quality of early estimates, and available funding[2]. The quality of early estimates, as referenced in that work, leads to selection of an "optimum" design, which is often later shown to be inefficient, ineffective, or impossible to manufacture, resulting in cost overruns and schedule slippage[3].

1.1.1 Fidelity-Forward Design and Its Challenges

When the models being used prove to be inadequate, there are two aspects that must be considered for improvement: whether the current model is understood thoroughly enough, or whether the current model is the correct model. A model that causes rework in one case might simply have been used for an assessment of behavior that was outside of its purview, or trusted region of application. This is a similar statement to the above quote by George Box, except in this case the takeaway is that while all models are wrong, each model was presumably developed with some application or applications in mind. As such, the model's usefulness is limited to the applications that the developer intended and venturing outside of that is not recommended. This is related to the other aspect of adequacy which is even more difficult to assess; whether the current model is correct for any application. Vague terms such as *correct* are not to be handled lightly, as they bring up long-debated philosophical concepts of what constitutes reality. These debates, as well as the work that has been done towards handling it in this context will be discussed in more detail in Chapter 3.

Instead of focusing on the inadequacies of early conceptual models, one of the philosophies that has been applied in an attempt to preemptively eliminate rework is sometimes referred to as "fidelity-forward" design[3]. This is often done by experience of seeing where the process went wrong in the past. An example of this is in the case of determining a manufacturing cost for an aerospace vehicle with new materials and/or manufacturing

processes, such as metallic joining techniques such as friction stir welding, or advances in the use of carbon composite laminates.

Typical low-fidelity models are based on interpolation, often with corrections, of historical data. Since the models being used have an abundance of data for more traditional concepts, the regressions are more readily believed. However, the estimated cost of manufacturing a composite structure is based on a small amount of data gathered presumably from early in the technology's development cycle when the costs per unit would be higher. To correct this, a more bottom-up cost estimation approach must be applied that accounts for the individual fabrication and assembly steps, providing more granularity, and as such, a better, more traceable estimate. This is referred to as Manufacturing Influenced Design (MInD), and has been applied to both aircraft and launch vehicle examples[4, 5].

This is all well and good for the cost estimation, but increased granularity in the cost model is representative of a more detailed expression of the manufacturing process in structural design, which is not available from the tools typically used during early design. As such, a higher-fidelity evaluation of the structure must be performed to provide the appropriate data for the new cost model[5]. This is a drastic increase in the complexity of structural evaluation typically performed at that point, which has provided a benefit, but has many drawbacks.

To execute the appropriate structural analysis and optimization, much more detail about both the vehicle and the environment must be provided. The additional information about the vehicle means that design decisions must be made that could previously have waited for a later point. Environmental conditions, specifically the loading scenarios, are much more detailed than those typically used, which requires another reevaluation of the model set.

In order to provide the newly required level of detailed loads, more information is typically needed regarding the aerodynamics of the vehicle in flight, be it an aircraft or launch vehicle. Additionally, the level of information related to the mission may have to be increased, and even, potentially, further disciplines. Not only has this single re-evaluation

trickled down into a complete reassessment of the entire multidisciplinary design space, all of the linkages between the disciplines must be verified and potentially reworked. On top of that, increasing the level of fidelity in every discipline could very easily cause an infeasible level of growth in the required effort to set up, execute, and process all of the models together. Because of this, the issues of model selection, development, combination, and interaction must be thoroughly explored.

This discussion brings up the question of why the models used in early conceptual designed were developed in the first place. Higher fidelity models that include the appropriate detail, physics, and scope require "major design commitments[6]." An example given by Lee is the design of a system to incorporate boundary layer ingestion in aircraft propulsion, where understanding the relevant flow characteristics requires specificying wing area, number of engines, and other characteristics of the design that may not have, as of yet, been specified.

In addition to design commitments, increasing fidelity can result in a dramatic increase in dimensionality. If design decisions have not yet been made, or the appropriate experiments performed, many of these dimensions must simply be defaulted. Setting many of the parameters of a model to default values increases the uncertainty in the model, which acts counter to the initial reason for bringing the model forward. As such, there is a time and a place where every model may be the most appropriate for some aspect of understanding a system. This makes it all the more difficult to justify what the most appropriate model is for a particular application.

1.1.2 Multifidelity Considerations

In an ideal circumstance, model selection would a purely scientific undertaking, conducted with quantitative rigor, through comparison of model data to truth data. However, if all of the data was available a priori to provide absolute certainty in model validity, modeling would no longer be necessary. In reality, even if a model can be proven to provide a

complete and exact representation of a system, such a model would likely take an incredibly long time to run. As such, fidelity is not the only factor to take into account.

When the desired level of fidelity is too costly on its own, a faster, lower fidelity, model can be used in conjunction with the high fidelity model to find a compromise position. Methods that incorporate multiple model evaluations into one source of knowledge can be called multifidelity, variable-fidelity, or variable-complexity[7].

An important reason for continuing to use lower-fidelity evaluations is that it has been shown that you can directly incorporate lower and higher fidelity models into the same design environment. One way that multiple fidelity models can be used together in optimization is by using the result of a lower fidelity model to improve the initial guess of the higher fidelity model. If the initial guess of higher fidelity optimization problem is improved, less iterations should be required to achieve convergence. However, this requires developing both models not only simultaneously, but in a directly connected manner. Additionally, it is not always applicable, since it only applied to optimization routines using similar-enough models.

As such, a more generally applicable multifidelity technique should allow for isolated model development, only combining data from each model a posteriori. This is often achieved through surrogate-enabled methods to increase efficiency through repeated evaluation of lower-fidelity models and evaluation of higher-fidelity models only as needed for uncertainty reduction. The magnitude of the response surface is therefore driven by the higher fidelity model, as it should be, while efficient interpolation is made possible by inclusion of more evaluations of the lower fidelity model.

An added benefit of multifidelity evaluation is that the design phases are no longer hard starting and stopping points. Put a different way, results are not generated using a single tool, used to make a decision, then thrown out. Where possible, the understanding of the system should blend directly from one phase of design into the next one: this is enabled by multifidelity analysis and optimization methods. Experimental data, when available, can

even be used to act as the highest level of fidelity in a multifidelity regression.

Some examples of multifidelity evaluations that have been proposed in the literature follow. Courrier, Boucard, and Soulier proposed a method for using partially converged solutions to improve efficiency in surrogate model building[7]. Many other works use forms of Multifidelity Gaussian Process Regression, also called Co-Kriging. This includes the work done by Le Gratiet[8], and is primarily built on the seminal work of Kennedy and O'Hagan: a Bayesian approach to multifidelity analysis and optimization using Gaussian process regression[9].

Some examples of applications of Co-Kriging to the field of aerospace engineering include the work done by Allaire and Willcox[10] and Ng and Willcox[11] towards a structural wing-sizing problem, Alexandrov et al. to the application of aerodynamic sizing of wings and airfoils in two and three-dimensions[12, 13], and the aerospace structural and aerodynamic applications in March's dissertation[14]. Throughout these works, the implemented multifidelity approaches typically achieved between a 50% and 90% reduction in computational evaluations of the higher fidelity models.

While multifidelity methods have been shown to provide a significant increase in efficiency, they do not simplify the model selection process. As discussed previously, it has always been difficult to find the appropriate model for a particular phase of design. As issues arise from an "intolerable loss of accuracy[15]," models are moved forward from their typical place in the process, changing the problem from development of a particular type of model, to selection and development of any potential model that can address the current problem definition. Incorporation of multifidelity considerations means that model selection now entails consideration of selection and development of any single model or combination of those models.

1.2 Overarching Research Question and Overview

Overall, what is being stated here is that higher-fidelity tools are needed to decrease the uncertainty inherent early in the design process. This is typically enabled by some combination of a more granular description of the problem and the more explicit incorporation of mathematical representations of physical constraints. However, this creates a drastic increase in the human and computational effort required to perform an evaluation, which does not lend itself to the quick turnaround and wide-reaching trade studies usually undertaken in early design.

One way of aiding in the efficiency of the generation of results when higher-fidelity models are used is to use them in combination with lower-fidelity models using multifidelity techniques. While this has the potential to increase the efficiency of the process, reducing the turnaround time, it has the opposite effect on the degrees of freedom in model selection. It is troublesome, therefore, that now models need to be developed and selected at various levels of fidelity, when it is still difficult to select and develop a model at a single level of fidelity. This leads to the motivating question of this dissertation:

Overarching Research Question *How can a model or models of the correct fidelity and efficiency be developed and selected to be used in a multidisciplinary, multifidelity design environment?*

From this research question, there are three main terms that require further discussion: *modeling*, *fidelity*, and what is meant by *correct* in this context. Chapter 2 consists of a more thorough discussion of modeling, the types of models, and why models are used. Chapter 3 discusses model credibility and understanding, the associated terms, and some of the methods used to quantify model correctness. This includes the discussion of validation, verification, calibration, accreditation, uncertainty quantification and sensitivity analysis. The discussion of fidelity and its associated terminology is outlined in the Chapter 4. A review of the descriptions of fidelity is presented as well as the development of a framework

for the description of fidelity in the context of this work.

Following that, methods are developed in Chapter 5 to understand and score the fidelity and efficiency of modeling options. These methods are developed using a set of four notional models and a set of fifteen finite element model representations of an I-beam. The methods are initially explored using the notional set, as the relative fidelity and efficiency can be varied to verify that the methods capture the intended effects. This begins by using the developed fidelity framework to allow experts to rank models, which is then used to calculate the probability that each model is the highest fidelity model in the multifidelity set. Then, a method is developed to leverage available data through comparative data analysis; this allows for troublshooting, initial down-selection, and adjusting the fidelity probabilities based on model data instead of just expert opinion. This method differs from other data-infused methods for calculating the probability of highest fidelity in the literature since it leverages model data even in the absence of validation data. The I-beam model set is primarily used to develop the comparative data analysis methods, as it incorporates different representations of a physical system while remaining geometrically and computationally simple, easing the development and processing of the associated cases.

Once the model fidelity probability methods are in place, the cost, or relative efficiency, of the models is assessed, and scoring methods are generated to compare single and multi-model options. This is what enables model selection, since, as discussed previously, model selection is dependent not just on fidelity, but on efficiency. Going to the highest fidelity possible is overkill at early points in the design process, and an understanding of the time-related costs help to make the decision-maker more aware of when cost is the limiting factor. In the case where the model of the desired fidelity is too expensive, it can be compared in terms of fidelity and efficiency to the other model options as well as multifidelity combinations. The relative scores can be used to justify a modeling path forward.

Chapter 6 describes the application of the developed methods as a full decision-making framework, examining a more realistic use case: a trade study examining the change in esti-

mated wing weight as the wing aspect ratio of the NASA Common Research Model (CRM) aircraft is deviated from its baseline value. The decision-making framework involves defining the problem of interest, enumerating the potential modeling options, making an initial assessment of fidelity based on the fidelity descriptive framework developed herein, adjusting the understanding of fidelity based on available model data, understanding model cost and effiency, and comparing single and multi-model combinations. The informed decision enabled by this framework is inherently dependent on the current point in the design process, as fidelity and efficiency limits can be imposed on the non-dominated single and multi-model options to find the current most appropriate modeling options. From there, the user can begin to understand what future cases need to be run based on the down-selected model or models, their relative efficiency, and which design points have already been evaluated. As models are further troubleshooted, verified, validated, and evaluated, the user can iterate back to update the understanding of fidelity and efficiency, adjust the requirements to a different point in the design process, and reevaluate the quality of the model selection.

CHAPTER 2
MODELING IN ENGINEERING

The first primary topic of discussion is the development and categorization of models. In the design of a system, there needs to be a way to evaluate aspects of that system. That evaluation provides an insight into the behavior of the system, typically through a quantitative representation, that can allow for design decisions to be made. This is typically done through models, as interacting with the actual system is difficult and expensive even in the case where it already exists. Section 2.1 describes in more depth the ways that a system can be studied, what a model can represent, and a first pass at how it can be represented. The next section, 2.2, gives insight into the way that models have been categorized and described in literature. This is intended to develop the understanding of the types of models in order to further discuss model credibility and fidelity in later sections and chapters.

The basic definition of a model is a representation of something else[16, p. 122]. Other, more specific definitions of **model** have been given.

[16] A representation of an event and/or things that are real (a case study) or contrived (a use-case), a representation of an actual system, or something used in lieu of the real thing to better understand a certain aspect about that thing[p. 5]

[17] A device which is so related to a physical system that observations on the model may be used to predict accurately the performance of the physical system in the desired respect[p. 57]

[18] Mathematical or logical relationships that are used to try to gain some understanding of how the corresponding system behaves[p. 1]

[19] A representation of a system or a part of a system in a mathematical or physical form suitable for demonstrating the way the system or operation behaves or may

be considered to behave[p. 107] *or* A qualitative or quantitative representation of a process or endeavor that shows the effects of those factors which are significant for the purposes being considered[p. 108]

[20] An idealization of part of the real world that aids in the analysis of a problem

2.1 Purpose of Modeling

Before any modeling and simulation is performed, there must be a need. There must be some aspect of the real world that that requires investigation, either a single aspect or in its entirety. This can be a single entity such as a vehicle or a system of entities together. The subject of interest in modeling and simulation has multiple names and, as implied by the definitions above, can be almost anything. These can be physical systems or processes, or the components or subsystems of those systems and processes, which are in turn smaller systems. Sometimes these levels have more specific names, an example of which is shown in Table 2.1. The system being referred to by a model is called the *prototype*, *simuland*, or just the system[16, 17, 18, 19, 20].

Table 2.1: Parts of a System[21]

Level	Specific Name
System	Launch Vehicle
Subsystem	Propulsion
Element	Liquid Engine
Component	Turbopump
Part	Turbine Blade

Figure 2.1 shows the ways to study a system as described by Law and Kelton[18]. What follows is a further detailed discussion of these comparisons.

Experiment with the actual system vs. Experiment with a model of the system :

It is preferable when possible to manipulate the actual system, operating it under different conditions, and observing the effects. Working with the actual system removes any question

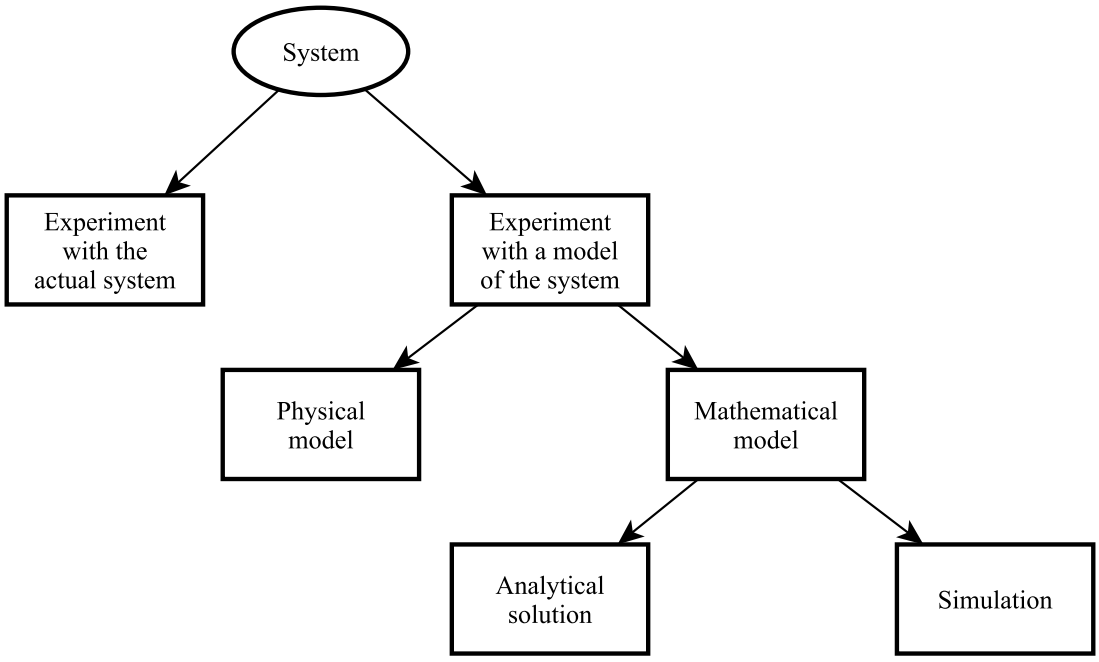


Figure 2.1: Ways to Study a System.[18]

to the relevance of the study. However, relying on manipulation of the actual system is a poor choice for a variety of reasons: physical experiments are too costly, manipulation could disrupt or damage the system, and especially in some circumstances, observing the system changes its behavior. This last point, generally referred to as the observer effect, can manifest in a few different ways depending on the exact scenario.

This is related to the uncertainty principle proposed by Heisenberg in quantum mechanics[22]. A change in a person's behavior when they know they are under observation is called the Hawthorne effect[23]. Similarly, the placebo effect occurs when someone presumes an effect should take place and a physiological reaction occurs without a pharmaceutical enabler[24]. The observer effect also comes into play in physical engineering experiments, e.g. inserting an instrument into a flow to measure its characteristics disrupts the flow field.

This type of effect even occurs in computer systems. Assessing computer systems or code can change the behavior of what is being observed by further taxing the system or

reassigning how processes are handled. This makes debugging or troubleshooting more difficult[25].

Additionally, and of significance in the case of this work, investigation based on manipulation of the physical system is often impossible, as the system being designed does not yet exist. Due to this, the manipulation would involve building multiple varied instances of the product which, for a large aerospace vehicle, is far too labor and cost-intensive to be feasible. As such it is typically necessary to develop a model as a stand-in or surrogate for the real system. The primary struggle with modeling lies in determining whether the model accurately represents the system under the circumstances of interest. This is the discussion of model *validity* which will be covered in greater detail in Chapter 3.

Physical model vs. Mathematical model :

Given that manipulation of the actual system is infeasible and a model must be developed, there are still two primary types of models. Sometimes manipulation of a scaled representation of a system can be useful. One of the primary examples where a physical model can be of help is in wind tunnel testing, whether the object of interest be a building, automobile, aircraft, launch vehicle, etc. One of the main enablers of this type of comparison is dimensional analysis[17]. Specifically pertinent to aerodynamics, the use of dimensionless similarity parameters, such as Reynold's number and Mach number, allow for a geometrically scaled model to generate a similar flow to the full object[26]. However, physical model tests still have some of the limitations of actual system tests. Experimenting with physical models is still costly and time and labor intensive, and the observer effect must still be taken into account. However, as with experiments on the actual system, the realism answers some of the questions of validity. The monetary expense and time required for physical testing, as well as increased computational power over time have led to an ever-increasing interest in mathematical representations of the systems of interest to supplement or supplant physical testing where possible.

Analytical Solution vs. Simulation :

Once a mathematical model has been developed, Law and Kelton describe the implementation of that model as either analytical or a simulation. Analytical in this context refers to a mathematical model that works in a straightforward manner to arrive at an exact, repeatable answer. Another term for this is a closed-form solution. Analytical solutions can be simple enough to work through with pencil and paper, or in the instance of inverting a large nonsparse matrix, they can require vast computing resources. The important point is that the answer is known in principle.

A simulation model solution is required when the system is highly complex, which leads to a complex mathematical approximation. In these cases, the model itself must be exercised, simulating the conditions of interest to determine the relationship between the inputs and the outputs, as they are nontrivial. This work is predominantly focused on mathematical models used in the design of products; most notably aerospace vehicles. In this investigation, it is not as important whether the model can be solved analytically or via simulation, though each presents different challenges. Typically, analytical models provide answers more quickly, but are representing the simuland as more simple than a simulation model would.

Deciding how to model a system is a difficult challenge. In 1950, Murphy[17, p. 57] spoke about how mathematical models are developed and used in engineering applications. He states that there are three situations that typically arise:

1. Direct application of well-known laws based on equilibrium or other conditions of state

2. Situations where number of variables or the complexity of the situation make the application of the usual analytical procedures tedious and may lead to a mathematically cumbersome solution

3. Problems where the general laws governing the behavior of the system are unknown

and analytical prodedures have not yet been developed

Whether or not a general formula can be generated from well-known laws, a relationship between design variables is needed. The applicable range for the variables is changed depending on the generality of the relationship. As well as the variable ranges, a model may be able to predict an entire scenario or maybe just one aspect.

Oftentimes models with narrower ranges or single-characteristic prediction are able to provide results more quickly and cheaply, which is an important characteristic in model selection. In other words, "more than one type of model may be useful in predicting the behavior of a given prototype[17]." This brings to the forefront an important characteristic about modeling; any given number of paths can be taken to arrive at an equivalent result. Building a specific model falls somewhere between an art and a science of applying a quantifiable relationship to represent a complex system. The art comes in because on top of that, the result could be arrived at in a variety of different ways. As such, the type of model needs to be described, and the way that validity is affected for that model needs to be assessed, both quantitatively and qualitatively. The next section speaks to the first of those points: How can the type of model being developed be categorized?

2.2 Types of Models

Models have been divided up into categories by many authors based on a variety of characteristics. Murphy[17, p. 61] divided models into four categories:

1. **True Models :** All significant characeristics are reproduced to scale and restrictions introduced by the design conditions are satisfied

2. **Adequate Models :** Accurate predictions of one characteristic may be made, but predictions of other characteristics are not necessarily accurate

3. **Distorted Models :** Design condition is violated sufficiently to require correction of the prediction equation, which usually pertains to different length scales within a

single model

4. **Dissimilar Models :** Bears no resemblance to physical system (what Murphy refers to as the *prototype*), but provides accurate predictions of behavior through analogies

An observation from how these categories are developed is that they predominantly speak to the relationship between two aspects:

1. The amount of detail present/resemblance of model to prototype
2. Amount that the relevant physics have been included, simplified, or distorted

Another important perspective that can be established from this is that the relationship between these two aspects and accuracy is not simple. Dissimilar models can provide accurate predictions despite bearing no resemblance to the physical system. Additionally, distorted models can provide accurate results despite an altered perception of real physics. These aspects will come up again later in the discussion of fidelity in Chapter 4.

Alber[27] divides models into four different categories:

1. Best Guess
2. Back of the Envelope
3. Complex Math Model
4. Large Numerical Simulations

Best guess approximations, while they have their purpose, are too rough and subjective to discuss with great rigor. The purpose of Alber's text is to develop the "art" of producing "back of the envelope," or rough approximations based on basic understanding of a problem to determine the order of magnitude of interest. This is a useful exercise to understand the important aspects of the problem, make use of limited resources, and find a starting point for further work. However, the models developed using these rough approximations

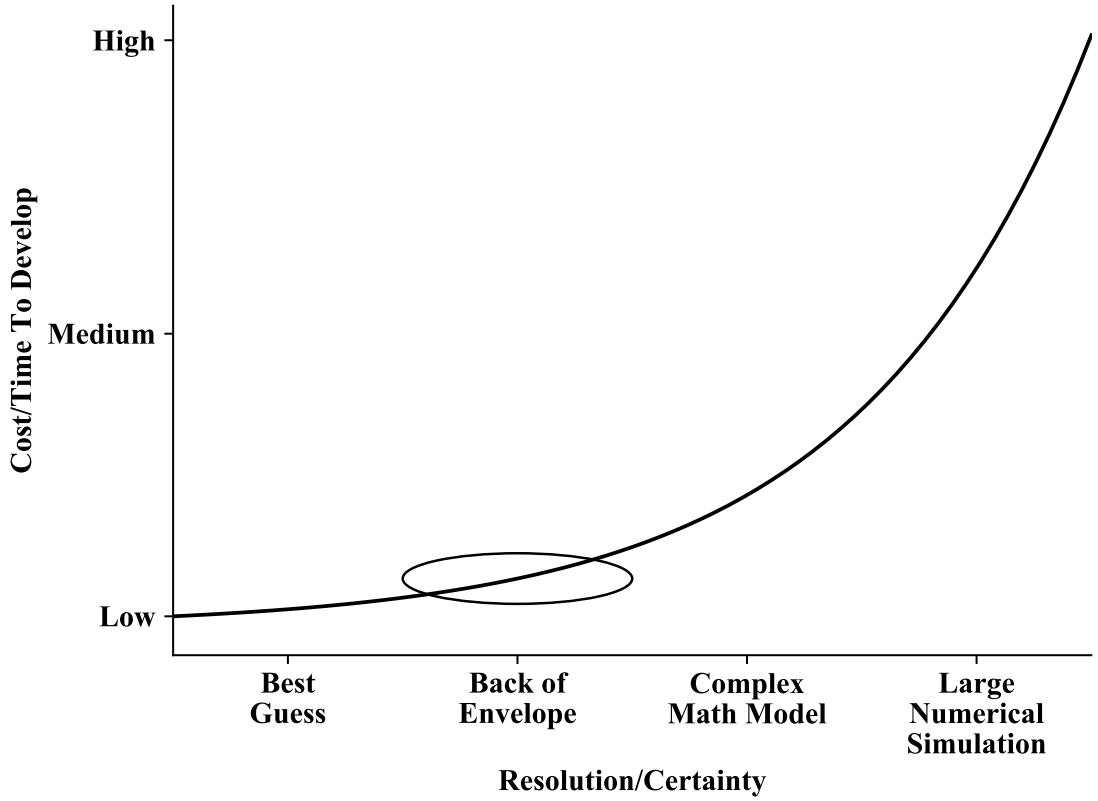


Figure 2.2: Relative Cost/Time to Develop an Answer to a Given Problem vs Resolution or Certainty of the Resulting Solution. Circle Denotes Range of Techniques Covered in Alber's Book.[27, p. 5]

also include broad-sweeping assumptions, potentially leading to gross inaccuracies. The latter two categories denote analytical or numerical methods that require computational processing. While general, the categories do describe differences between models related to accuracy and required effort; models that more accurately represent the incorporated physics, have less broad-sweeping assumptions, and a more detailed description of the system, generally produce results that are more certain. The caveat is that these models can require significantly more time and effort to develop and evaluate, as notionally represented in Figure 2.2.

Chestnut[19, p. 114] discusses models, both qualitative and quantitative, that are used in systems engineering. The categories of models as well as a variety of purposes for those

models are described in Table 2.2.

Table 2.2: Categories and Purposes of Models as Desribed by Chestnut[19]

Category	Purpose
Overall Process	Schematic diagram showing major elements of hardware Functional diagram showing operation of system Procedural description showing time sequence of operations
Performance	Accuracy Speed of response Material balance Energy balance
Time	Overall schedule Detailed schedule Program Evaluation Review Technique (PERT)
Reliability	Parts Equipment System
Cost	Parts Cash flow

He continues [pp. 125-127] by describing a system for classification of models through three categories and a series of examples for each:

1. Resemblance to reality

   - *Isomorphic :* Model which includes *all* of the operational features of the real situation
   - *Homomorphic :* The more common case where related variables are grouped to represent general effects instead of a detailed representation of all effects; this type of model should presumably be based on physical laws or grounded techniques for describing the behavior of the pertinent phenomena

2. Type of models: structuring characteristics

   - *Iconic Models :* Models which "look like" the subject but may be scaled to a more manageable size, e.g. globes, wind tunnel test articles, and so forth

- *Analog Models :* "Characterized by the use of a convenient transformation of one set of properties for another set of properties in accordance with specified rules," such as converting a manufacturing process to a flowchart or equivalent circuits
- *Symbolic Models :* Logical and mathematical representation of behavior of the system or components of the system at hand

3. Method of solution

   - *Solution by analytic methods :* Explicit determination of a result from a straightforward application of mathematical techniques which are typically applied to relatively simple cases and can be tested against a real scenario
   - *Solution by numerical or deterministic methods :* Typically an iterative process involving computer solutions of models for testing specific states to determine the critical conditions
   - *Solution by Monte Carlo methods :* Another iterative process involving computer solutions, but one that involves testing states of the model to determine the probabilistic properties of the system instead of the deterministic behavior

Aside from describing how to study a system, Law and Kelton[18] discuss that the specifics of their "simulation models" are described by categorization along three dimensions:

1. *Static vs. Dynamic :* This dimension is summed up fairly succinctly as whether the model is time-dependent or not. A static model is often what is used for the design of vehicle structures since the structure remains roughly the same over time (barring damage) To the contrary, the important behavior in system design is how it evolves over time.

2. *Deterministic vs. Stochastic :* If the system contains components with inherent randomness, then they are stochastic. However, if the result is consistent, the model is deterministic. Put a different way, once the inputs and the model relationships have been put in place, the output is "determined." There are pros and cons to this characterization: reality always contains some amount of randomness, randomness in computer modeling is difficult, and while modeling a stochastic system as such is accurate, the results are just a single estimate of behavior, e.g. repeated trials and post-processing are necessary.

3. *Continuous vs. Discrete :* The type of change of the "state variables" describe the continuity of the system. In a discrete model the system attributes change instantaneously, or "at only a *countable* number of points in time." A continuous model describes how the system got from one state to the other, e.g. a description of traffic flow as a whole instead of by the individual car.

The discrete-event simulation models of interest to Law and Kelton are typically discrete, dynamic, and stochastic. However, the physics-based vehicle design models of interest here are typically static and deterministic. Instead of quantifying behavior over time, snapshots of scenarios the vehicle might go through are used as constraining cases. Randomness, when it is included, is typically applied via safety factors, knockdown factors, and other corrections based on probabilistic assessment of real-world behavior.

Dieter[20, pp. 247-249] describes models more in terms of their purpose for design. Models, he states, are either descriptive or predictive. A *descriptive* model is a qualitative decomposition, but does not describe the behavior in terms of extent. A *predictive* model, on the other hand, provides a prediction of the system's behavior in order to understand the type and extent of its qualities. Models can also be classified in three other ways:

1. Static or Dynamic

2. Deterministic or Probabilistic

3. Iconic-Analog-Symbolic

   - *Iconic :* "Models that look like the real thing." These are often intended more to represent the entities than the phenomena. They are typically either two-dimensional (e.g. maps and drawings) or three-dimensional (e.g. physical models, CAD) and can be further categorized
     - *Proof of concept model :* "Minimally operative" model to represent basic principle of design concept. Sometimes referred to as a "string and chewing gum" model.
     - *Scale model :* Dimensionally resized typically non-operational portrayal of physical world, often used for discussion and visualization of basic concept or interferences.
     - *Experimental Model :* Model built to represent the function but often not the aesthetic of the design concept in order to perform testing and design modification.
     - *Prototype Model :* Full-scale working model of the design. Similar if not identical to final product in technical and visual aspects, but typically built by hand as a single instance. This is the final selling point for a product to go to full production.
   - *Analog :* As the name states, these types of models are analogous to the actual system. This means that they are able to predict the system behavior but may look nothing like the actual system. Process flow charts are one example of an analog model that is given.
   - *Symbolic :* This is where mathematical models predominantly reside in this categorization. The relationships between input and output parameters represented in a mathematical equation is a type of symbolic model. Symbolic models lead to quantitative results via analytical, mathematical, and logical ex-

ploration. They can also be further divided into two categories similar to the scenarios put forth by Murphy above[17, p. 57]:

- *Theoretical models :* Based on "established and universally accepted laws of nature."
- *Empirical models :* Best approximate mathematical representations based on existing experimental data.

As previously discussed, *simulation* is the process of exercising a complex model once it has been developed. This aides in the understanding of the relationship between inputs and outputs and analogously some semblance of the behavior of the real system under investigation.

These categorical descriptions attempt to be general, but aspects of them are specialized into two categories: descriptions of the time-dependent behavior of entities, such as in discrete event simulations, or the physics-based description of the system for the purpose of design and analysis. Put another way, the terminology used varies based on whether the model is describing how entities behave from a detection/decision/reaction standpoint or whether the model describes how the aspects of the system react due to the behavior of materials, flows, etc. There is obviously much overlap between these two categories, but the distinction makes it such that certain aspects of a categorical decomposition are not generally applicable to all possible models.

Similarly to Alber's categorization of models, an altered representation of generic model description is presented, split into four categories.

1. Regression of Existing Data
2. Simple Analytical Approximations (e.g. Back of the Envelope)
3. Analytical Models
4. Discretized Numerical Methods (e.g. FEA, CFD)

It should be noted that these categories are not necessarily in order by increasing accuracy. *Simple analytical approximations* are "quick and dirty," which is good to get a fast approximation. In the presence of existing trusted data, fast approximations using regressive techniques can be very accurate. The downside in that case is that you are limited to the scope and assumptions of the training data set. *Analytical techniques* can be very true to physical behavior but are also very limited by the requirement of explicit mathematical solutions. Deriving a precise formulation gets more difficult as complexity increases. The benefit of *discretized numerical methods* is not necessarily in accuracy as often these models require calibration (e.g. dynamic finite element models and most computation fluid dynamic models). They are, however, quite useful for their flexibility. This flexibility comes at the cost of a drastic increase in the amount of required data and runtime that scales relative to the discretization and underlying equations.

Models are useful tools for thinking, communicating, predicting, controlling, training, exploring, and designing. The increase in computing power has enabled the implementation of more and more advanced models that cannot be solved using classic analytic methods. Modeling is and should be used "early and often" due to its capability to help sell a product or concept by predicting behavior as well as providing a visual representation. Moreover, mathematical models enable exploration of systems at a lower cost, in less time, and when no physical instance yet exists. Throughout an engineers work, a "menu of models" are developed that are used throughout the thought process[20, pp. 249-250]. However, the benefits of mathematical models must be weighed against their validity in describing a given scenario when selecting the appropriate model from this "menu." The following chapter discusses how the accuracy and uncertainty in a model is typically assessed.

CHAPTER 3
MODEL CREDIBILITY

The next main point of discussion is the use of the word *correct* with respect to modeling. Determination of correctness or accuracy in modeling and simulation is related to a number of terms that will be discussed here, such as verifiability, validity, credibility, accreditation, and the field of uncertainty quantification. Another way to put it is the determination of whether the model is *valid*[18]. This is especially important as design decisions are to be made based on the results of the selected models. If the results are not trusted then the decisions made based on them cannot be trusted.

To frame the problem, a discussion of the procceses required to believe that a model is adequate is necessary. At the end of the day, whether or not a model is to be believed comes down to a sales pitch of whether all parties involved believe the results: managers, clients, analysts[18]. Since this is a technical field, credibility can be understood, to some extent, in objective terms. This is why there are standards in various fields of study to be able to quantify confidence in terms that an informed individual can understand. These and other requirements attempt to guarantee that sufficiently thorough analysis was performed. It is ultimately always a subjective matter as to how much is required to instill confidence. Because of this, much research has been done to explicitly define the required steps.

This section provides some background into the basics of proving that a model makes appropriate predictions for a given application. The provided background aims to denote the difficulty in proving that a model is correct, especially in the case of early design when available data is sparse. It also serves to denote the ways that uncertainties have been described in order to facilitate further discussion of model fidelity in Chapter 4.

Some preliminary definitions must be provided here. The context of this work is to provide insight in the development and understanding of models typically used for design

and analysis of products such as aerospace vehicles and their systems. These types of models, incorporating physical laws through scientific models, are implemented into computer software to allow for modern computational power to aide in the design and analysis process. Once implemented, these models often fall into what is called computational science and engineering (CS&E). Once a model has been implemented into software that can be executed, those software systems are called *codes*. The product of the execution of these codes can be called *outputs*, *responses*, or a *calculation*[28].

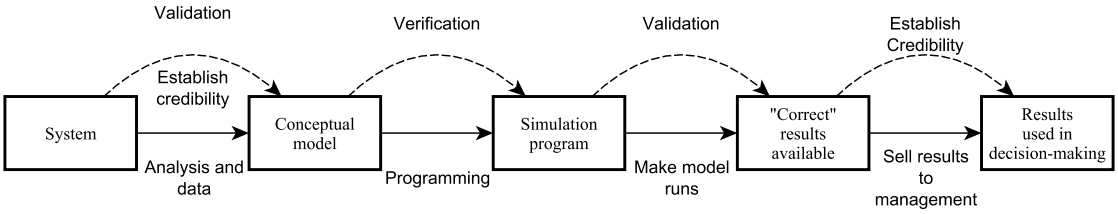


Figure 3.1: Timing and Relationships of Validation, Verification, and Establishing Credibility[18]

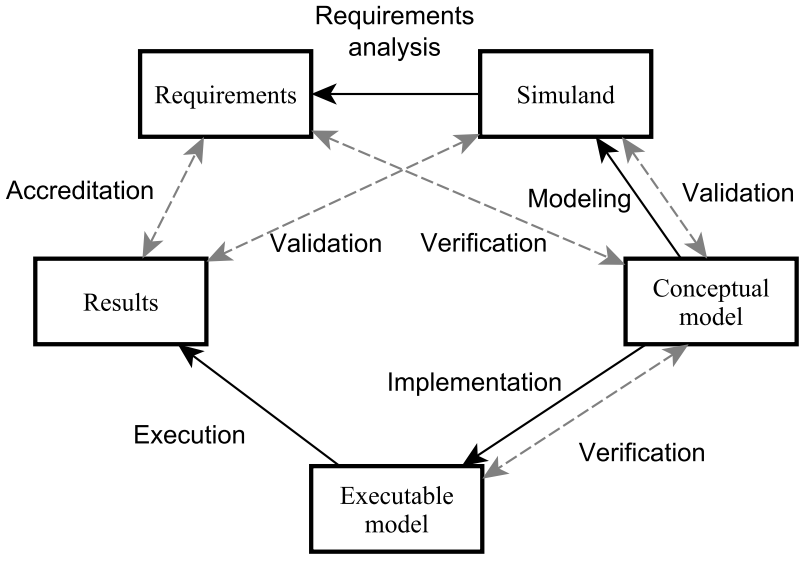


Figure 3.2: Comparisons in Verification, Validation, and Accreditation[16]

3.1 Benchmark

One of the required items to perform most of an evaluation of a model or code is a benchmark. The definition given by Trucano et al. follows:

- A choice of information that is believed to be accurate or true for use in verification, validation, or calibration, one or more methods of comparing this information with computational results, and logical procedures for drawing conclusions from these comparisons[28].

This is a very interesting definition as a benchmark might typically just be thought of as a model. However, Trucano et. al. make sure to note that the model is not useful as a benchmark unless it is accompanied by the methods of comparison and drawing conclusions from the data from both the benchmark and that generated by the current model or code. It is of further note that the benchmark is itself some abstraction of reality. This could include an "analytic mathematical solution" based on fundamental physics and compared to a real scenario or an experiment that is deemed to have low enough error for use in this purpose. The choice of benchmark is also inherently dependent on the intended application, so the process of justifying a benchmark is similar to the procedure for justifying the comparison of a model to a particular benchmark.

3.2 Verification and Validation

Once a model has been implemented, that implementation, or *code* must be verified and validated. These are two distinct processes but are usually grouped together in discussion due to their necessity and interrelationships. A succinct description of these processes states that verification is "solving the equations right" where validation is "solving the right equations[29]." The difference between these two is subtle but important, and understanding the exact meaning and process involved with each has been the focus of much research.

3.2.1 Verification

The first step that must be performed upon generating a code is verification. The process of verification is related to uncertainty quantification and has an influence on validation, calibration, and whatever other processes follow it. There are multiple definitions put forth for verification:

- (US Department of Energy's Advanced Simulation and Computing program, ASC) The process of confirming that a computer code correctly implements the algorithms that were intended[28]
- (DoD) The process of determining that a model, simulation, or federation of models and simulations implementations and their associated data accurately represents the developer's conceptual description and specifications[30]
- (American Institute of Aeronautics and Astronautics, AIAA) The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model[28]

Verification is a necessary process to assure that as much spurious behavior as possible is weeded out prior to checking the code for accuracy to reality in any way. If verification fails in some way then any decisions made on predictions after that could be based on the skewed behavior of the model instead of its prediction of reality. One primary purpose of verification is to prove that the chosen solution algorithms and implementations thereof are correct. Assessment of applicability of a code must also be done in this phase. It must be assessed whether the perspective model can provide the type of response needed for the current application at all, regardless of its accuracy in doing so. In Figure 3.1 validation is described as the process of establishing credibility through analysis and data that the system is being represented by the conceptual model. This is also shown in Figure 3.2 as verification being the process of assuring that the conceptual model meets the current

set of requirements. Many of the techniques used in verification are related to software engineering and systems engineering[16].

3.2.2 Validation

Similarly to verification, there are multiple definitions put forth for validation:

- (ASC) The process of confirming that the predictions of a code adequately represent measured physical phenomena[28]
- (DoD) The process of determining the degree to which a model, simulation, or federation of models and simulations, and their associated data are accurate representations of the real world from the perspective of the intended use(s)[30]
- (AIAA) The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model[28]

Validation is the general process for determining that the code provides results that accurately represent the physical response or responses of interest. This is important as it is predominantly the aspect to which fidelity typically refers. Model fidelity relates to the validation of the model assuming that sufficient verification has been performed and appropriate calibration will be done given a referent. In Figure 3.2 validation is shown as the comparison between the results and the simuland.

To be specific, validation is the assessment of the behavior of that response. This is related to the discussion of accuracy and precision. In the case of archery, accuracy would be hitting near to the center of the target, but precision refers to the ability to strike the target repeatedly in the same region. In the case of a numerical model precision is typically the preferred attribute. If the model can predict the trends of the behavior correctly and repeatedly then the parameters can be adjusted to align those trends to the "center of the target" to refer back to the archery example. This process of adjusting to an accurate benchmark is called calibration.

3.3 Calibration

- (Trucano et al.) To adjust a set of code input parameters associated with one or more calculations so that the resulting agreement of the code calculations with a chosen and fixed set of experimental data is maximized[28]

As mentioned above, calibration relates to the adjustment of parameters to "best match" the code's calculation to a trusted data set[31]. There has been much work done on the process of calibration in the literature. One of the most commonly used traditional methods of calibration uses a least-square regression model[28]. Much of the more recent work done has been based on the Bayesian calibration approach put forth by Kennedy and O'Hagan[32]. Some examples of its usage include Sankararaman and Magadevan[33], Lacaze and Missoum[31], and Heo et al.[34] among many others.

3.4 Accreditation

Whereas validation can be described as "did I build the thing right," and validation is "did I build the right thing," accreditation can be described as "is it believable enough to be used[35]?"

- (DoD) The official certification that a model, simulation, or federation of models and simulations and its associated data are acceptable for use for a specific purpose[30]

Table 3.1 is adapted from the work of Balci and describes the general types of errors that can occur in the process of selecting a model for use. *Type I Error* is also called the *Model Builder's Risk*. A Type I Error occurs when a relevant and valid model is developed, but then not given accreditation and put into use. For one reason or another, not enough validation is performed to assure the proper parties that the model is correct, so the model is not chosen. The model provides credibility but any potential utility is lost. The time and money that went into developing, verifying, and validating it also go to waste. The risk

belongs to the builder of the model because any other negative effects are simply unfulfilled benefits that could have been garnered from its use. This type of scenario predominantly falls to the subjectivity of the situation, as it is caused by some combination of the developer not selling his product adequately enough or the potential accreditor being unnecessarily skeptical or cautious.

Type II Error can be referred to as the *Model User's Risk*. This error occurs when a model that is invalid is selected. Validation is performed incorrectly in some way but sufficiently enough to erroneously persuade the accreditation process to continue. When this occurs, improper results are believed and used for decision-making. This is the standard risk that the validation process attempts to plan against, as its occurrence can lead to disastrous consequences if left unchecked.

Type III Error, or *Model Accreditor's Risk*, must not be confused with Model User's Risk. A Type III error occurs when a model is used outside of its range of validity. This is different from type II error because it refers to a model that is valid somewhere. The problem is that it is used for an application where it is not relevant. As stated by Petty, "Type III errors are distressingly common; models that are successfully used for their original applications can acquire an unjustified reputation for broad validity, tempting project managers eager to reduce costs by leveraging past investments to use the models inappropriately."

The main trouble with Type III errors is that they are due to subjectivity, specifically familiarity, similarly to Type I errors, but they have all the potentially catastrophic downsides of Type II errors. The only two correct scenarios are when a relevant and valid model is used, or a model is not used due to irrelevance or invalidity. This is why there is such an argument for understanding your model based on more than just the comparison to reality.

The description of the model is useful for the understanding of what it does and what characteristics it contains. The relevant ranges and assumptions that went into its development have to be maintained and any potential user must be informed lest decisions be made using data of no relevance. Complex models will almost always have some range within

Table 3.1: Verification and Validation Error[16, p. 134]

	Model valid	Model not valid	Model not relevant
Results accepted, model used	Correct	**Type II Error**	**Type III Error**
Results not accepted, model not used	**Type I Error**	Correct	Correct

which their predictions are appropriately accurate and ranges where they aren't, sometimes referred to as the *bounds of validity*[16, p. 135]. This is by no means a justification for disqualifying its predictive capability outright. It is instead just a consideration in the selection process. One of the contradictions in the process of model selection is that it would, at face value, appear to a developer that being explicit about the limitations of the product would not aid in the sales pitch. People do not like to think in the negative and focus on what can't be done instead of what can. However, not being forthright regarding constraints to applicability is a dangerous proposition for the user.

3.5 Uncertainty

The discussion of uncertainty is a philosophical debate, focused on defining reality, its inherently random nature, and human perception. Moving from that, it pertains to whether an abstracted form is the reality itself and what those differences are. This discussion has been particularly embroiled in the field of quantum theory related to the probabilistic nature of reality, a point especially debated between Albert Einstein and Niels Bohr[36]. Generally speaking, at every level an abstraction occurred uncertainties were compounded; any idealizations, mathematical representations, or deterministic computations make the problem tractable but increase the level of justification required to prove that a model is valid.

The field of uncertainty quantification is a rapidly changing landscape in recent years,

and much work has been done to evolve the methods used, make them easier to implement, and work to spread their use in the industry. The ability to justify to a company that the computer and person time required to evaluate uncertainty justifies the amount of information gleaned from the exercise is a lofty but important goal. To make this justification, the problem must be developed not only numerically but qualitatively, so that the topics can be discussed in a way that is manageable to all parties.

Roy and Oberkampf provide many sources for the work in risk assessment that has been done to categorize the types of uncertainty in computational analyses[37]. However, they make the distinction that any description should refrain from confusing the nature, essence, or essential type of uncertainty with the description of how or where it might occur in the process. The example provided is if randomness is one category and model form uncertainty is another. "A sound taxonomy would *only* categorize uncertainty types according to their fundamental essence, and then discuss how that essence could be embodied in different aspects of a simulation[37, p. 51]."

The risk assessment community, led by the nuclear reactor safety community has developed one of the most workable, effective, and widely accepted descriptions of uncertainty. This description breaks uncertainty down into two main concepts: aleatory and epistemic uncertainty. Aleatory uncertainty, which has also been called intrinsic, inherent, irreducible, or stochastic uncertainty or variability[38, 37]. The word aleatory derives from the Latin *alea*, which refers to the rolling of dice[39]. This is to denote that aleatory uncertainty has its basis in the inherent randomness of a situation.

The other category is epistemic uncertainty, which has also been called reducible uncertainty, subjective uncertainty, state-of-knowledge uncertainty, or simply uncertainty[38, 37]. The word epistemic comes from the Greek $\varepsilon\pi\iota\sigma\tau\eta\mu\eta$ (episteme), meaning knowledge[39]. This infers that the uncertainty that can be described as epistemic is due to a lack of knowledge. This can also include a lack of data.

The application of these definitions is not always straightforward, as some factors can

move from one category to the other as design decisions are made. One example of this is how as a material is selected, the stiffness of a structure goes from aleatory to epistemic uncertainty. However, this is still somewhat ambiguous as there is random variation in real materials due to processes such as fabrication and assembly.

As the definition describes, there will always be intrinsic or inherent aleatory uncertainties related to any phenomenon being modeled. This work will predominantly be focused on epistemic uncertainty and in some cases combinations of epistemic and aleatory uncertainty. A minor reason for this is that the representation, aggregation, and propagation of aleatory uncertainty is related to traditional probability theory and is a well established field[38]. The more important reason is that this work is focused on uncertainties related to models, specifically models used for design in the early phases. At that point in the design phase, there is a great deal of reducible or epistemic uncertainty in general. Additionally, most uncertainty related to model development and selection are forms of epistemic uncertainty[2].

3.5.1 Types of Uncertainty: Robertson Dissertation[40]

One literature review on the types of places where uncertainty present themselves can be found in Robertson's dissertation. That review is aimed at identifying uncertainties in the realm of space and launch vehicles, but includes an exploration of uncertainty in the fields of ecology, conservation biology, civil engineering, structural engineering, systems engineering, modeling and simulation, space architectures and systems, and the design of complex systems. Of particular importance to that application is the differentiation between endogenous and exogenous sources of uncertainty.

Endogeneous

Endogeneous sources of uncertainty are defined as follows:

- "Sources of uncertainty which can be traced back to sources with a program office's

control, traditionally from three factors

1. Assumptions about new technology performance
2. Assumptions in early design that are overly optimistic
3. Assumptions about subsystems that are typically omitted in early design"

Exogeneous

Exogeneous sources of uncertainty are defined as:

- "Sources of uncertainty whose origins can be traced to outside of a program development office, primarily referring to uncertainties in requirements"

While both of these types of uncertainties have significant impacts on programs under development by an organization, this work is predominantly concerned with the development and selection of models that would occur within an organization. As such, this would fall more under the category of endogenous uncertainty. Exogenous uncertainties from changes in requirements or political instability would come into the probabilistic risk assessments done using the models once they have been selected or developed, but the model itself is much more concerned with the assumptions that went into its development.

Continuing with Robertson's work[40], the taxonomic decomposition arrived at is shown in Figure 3.3. *Endogenous* uncertainties, being uncertainties that are epistemic and within the purview of the program's office, can be reduced by the appropriation and application of applicable resources. This process could be a risk mitigation program or it may just occur through the natural progression of the design process. The epistemic, endogenous uncertainties are further broken down into *phenomenological uncertainties*, *human errors*, and *design uncertainties*. Much of the taxonomy put forth by Robertson is adapted from Melchers[41] and Thunnissen[2].

Phenomenological uncertainty, as the name would suggest, refers to how phenomena are represented. These could be physical phenomena or otherwise, and a lack of knowl-

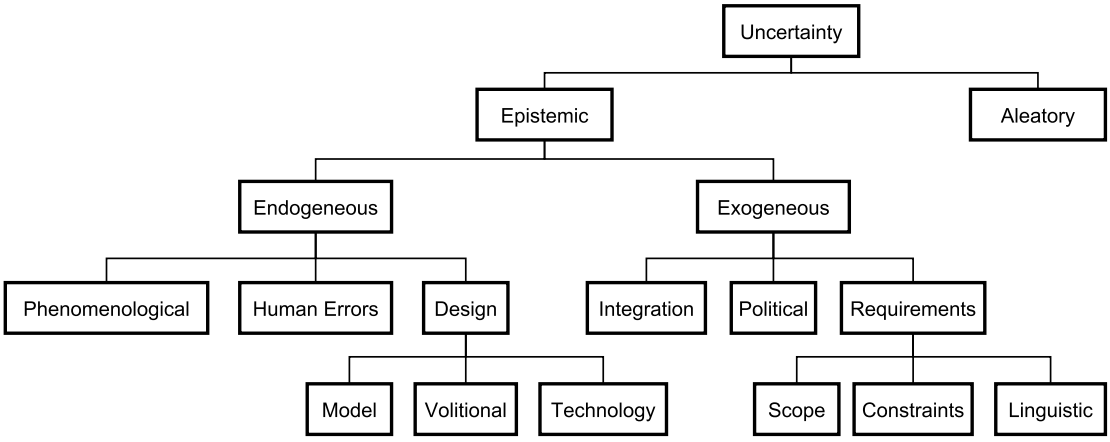


Figure 3.3: Taxonomy of Uncertainties in the Development of Space and Launch Vehicles[40]

edge of a given phenomenon comes across as an inability to recognize potential risks or opportunities that it could affect. These are typically referred to as "unknown unknowns" or "unimaginable" phenomenon[41] and are most appropriately referenced in the development of novel concepts. This is due to the fact that novel concepts are more likely to present behaviors that have not been foreseen, as they have not been widely used and observed. For more information on phenomenological uncertainty, see [42].

Human Errors are errors caused in the development by some mistake in the design process that goes unnoticed. This type of error in the design, manufacture, test, or operation of a system, can lead to anything from a small amount of rework to a loss of parts during operation. These are issues that can only be addressed through organizational and behavioral processes related to verification that are designed to minimize the ability of an issue to slip through.

Design Uncertainty is the largest of the three subcategories of endogenous uncertainties. Design uncertainties are due to a lack of knowledge in the design of the system in question. The way that this type of uncertainty presents itself is related to the current point in the design process. Design uncertainties are either due to assumptions made in previous design phases or they are due to uncertainty in design decisions that have yet to occur. This

category of uncertainty is further broken down into three subcategories: *model*, *volitional*, and *technological* uncertainties.

The first type of design uncertainty is *model uncertainty*. This type of uncertainty has in the past only referred to the epistemic uncertainty related to the fidelity of the analysis tools being used. However, Robertson makes an interesting note that is related to this work. Model uncertainty can also refer to the quality of the "fidelity level of the engineers' and designers' mental model of the system under development[40]." Put another way, the model uncertainty is related not only to the fidelity of the model itself, but to the quality of the data being passed to the model and the user's understanding of the model and the associated data. This is related to the concept of "garbage in, garbage out." Since a model representing reality always contains uncertainties and abstractions of that reality, then the information given to the model also has uncertainties and abstractions based on the perspective of the person using the model[43]. In this way, modeling uncertainty refers not to the fidelity of the model specifically, but of the overall fidelity of the implementation of said model.

The next subtype of design uncertainty, *volitional uncertainty*, is "uncertainty due to the decisions of actors within the design process of the system[40]." Where *model uncertainty* in this context refers to the uncertainty relative to the fidelity of a single model, *volitional uncertainty* is how that is expanded into the realm of multiple tools, models, and fidelities. The variation included here is caused by the decisions of individuals or program offices throughout the organization that is designing the product. Put more specifically he refers to two primary ways that volitional uncertainty can present itself.

The first way that volitional uncertainty manifests is through the future inclusions to the design process that add fidelity or refinement to the design. As detail is added to the design through the development and use of higher-fidelity tools, information about subsystems are added to the model information that was not previously included. If the subsystem was not included in the lower-fidelity evaluation it could be that it was assumed to not be of

great importance to that level of estimation. However, and what might be a cause of greater uncertainty, is if a new subsystem is added to the design through the process of refinement that wasn't there previously, implicitly or otherwise. This could potentially change the design, but the original method of approximation may be capable of representing the new subsystem, creating uncertainty between the different levels of modeling.

The second way that volitional uncertainty presents itself is due to future design decisions. As higher fidelity models are used and data is generated numerically or experimentally, more informed design decisions occur. Volitional uncertainty occurs when these design decisions contradict the previous understanding. One example of this is if the chosen design architecture doesn't meet the estimated level of performance, a change may be made. Additionally, higher-fidelity analyses could show that the design is infeasible or unmanufacturable, which would require some amount of redesign and rework. These issues fall under the category of volitional uncertainty.

The third category of design uncertainty is *technological uncertainty*, which is uncertainty caused by the inclusion of new technologies into the system being designed. New technologies are often included in designs for a number of reasons: seeking a step-change in performance such as the incorporation of carbon composite structures into the majority of aircraft and even launch vehicle structures, to overcome a challenge, technical or otherwise, or, as is the case with NASA, if part of the organizational motivation is to promote and aid in the development of cutting-edge technologies for the purpose of improvement our understanding and abilities as a society[44]. Due to the phase of development of the technology in question, situations may arise where the performance of a technology falls short of the optimistic projections set forth by the technological developer. Another situation could arise where a technology performed adequately but proved to be too difficult to fabricate, assemble, or maintain. A subsystem containing a new technology could also simply fail to work. Due to these types of shortcoming or some other type of failure to perform, a subsystem may have to revert to a more conventional alternative which can have

substantial effects on a design.

3.5.2 Types of Uncertainty: VUCA

A descriptive acronym that has gained wide usage in the business management and military strategic decision making communities is *VUCA*: Volatility, Uncertainty, Complexity, and Ambiguity.

Volatility describes when a good amount of information is known about a situation, but there is still a challenge in the fact that the behavior is "unexpected or unstable and may be of unknown duration[45]." It is related to the rate of change of the environment, which is unknown[46]. Examples of this include: price fluctuations that may occur after a natural disaster, the inability to predict exact weather conditions for a day-of-flight assessment of aerodynamic and thermal characteristics, and any other difficulty in predicting the exact operating environment[45]. These are purely aleatory uncertainties that can only be accounted for through the inclusion of margin, which could also be called slack. As with other aleatory uncertainties probabilistic analysis is required which is an area where much research has been done.

Uncertainty in this more specific context is the inability to determine the cause of a behavior despite information about the behavior[45]. Put another way, observing a situation does not necessarily provide you the insight to be able to predict the nature and effects of change[46]. A business related example of this is how the pending release of a competitor's product creates uncertainty in the market and the future of the industry[45]. Theses are unknowns about the intrinsic characteristics of the system and the environment but can negative affects can be hedged against using risk management and through the accumulation of additional data.

Complexity relates to the "interconnected parts and variables" inherent to most any true system. It can be difficult to understand the information about the environment either because of a lack of data or the inherently overwhelming nature of the factors at play[45]. Due

to the highly interdependent nature of the actors, a lack of traceability works against your understanding of the reasons for system behavior. This is a common occurrence in systems of systems where emergent behaviors appear due to the interdependencies of the individual entities. The volume of potential options muddles the view of the possible outcomes and presents an intrinsically difficult problem[46]. This is a situation of epistemic uncertainty where a better understanding of the individual contributors via disciplinary expertise, systems traceability, or other approaches can improve the understanding of the problem at hand.

Ambiguity is the appearance of "unknown unknowns" where the "causal relationships are unclear[45]." In military strategic planning this can refer to difficulties in gaining understanding due to cultural blindness, cognitive bias, or limited perspective[46]. This leads to the possibility that a situation may be interpreted in more than one way at a point when it is impossible to determine which of the paths is correct. This is a situation where little information is known and the prediction of behavior is poor, so the typical strategy is simply to performing experiments to test the system so that both prediction and knowledge can be improved[45].

3.5.3 Types of Uncertainty: Kennedy and O'Hagan[32]

This work is focused on the uncertainties that are caused by the model itself. Modeling-induced uncertainties provide a slightly different perspective for the decomposition of uncertainties. One such method of classifying uncertainty is put forth by Kennedy and O'Hagan to describe the various sources found in the use of computer models[32], specifically in the circumstances of Bayesian calibration of computer models. They decompose the issue into six categories:

- Parameter uncertainty
- Model inadequacy

- Residual variability
- Parametric variability
- Observation error
- Code uncertainty

Equation 3.1 shows the model used to represent the relationship between the observations, process, and computer model used in the Kennedy and O'Hagan Bayesian calibration method. In this instance, z_i represents the i^{th} observation, and the true process is represented by ζ. The true process is made up of η representing the computer model output, dependent on the design variable vector x_i and the calibration parameter vector θ, and the model inadequacy δ. The i^{th} observation error e_i is represented by a random variable.

$$z_i = \zeta(x_i) + e_i = \rho\eta(x_i, \theta) + \delta(x_i) + e_i \tag{3.1}$$

A more in depth description of the categories follows.

Parameter uncertainty is the uncertainty about the value a particular input parameter in a model that is not known and is therefore a reason to perform calibration. The values of inputs are intended to denote specific instances of the real world scenario. Because of this, the input may have a specific value for a particular application context, the same value for a range of instances of the scenario, or the same value over the full range of application for that model.

Model inadequacy is related to the discussion of every model being wrong to some extent. Even if the values of the inputs are known, there is still going to be some discrepancy between the model and a more trusted data set due to inherent random variability. Model inadequacy more specifically is defined as the difference between model response and the true *mean* value in the real world. This brings up the additional issue of knowing the true value of the process. In many instances, the calibration parameters of a model may

be related to "concrete physical meanings" that would seemingly denote a value based on the real world instance. However, if you fix these values early in the process, you are taking flexibility out of your model in its ability to predict the real world behavior. This is analogous to fixing or removing terms from a nonlinear regression model, limiting its ability to fit to the relevant data. This can become especially evident based on the results of a sensitivity analysis (discussed in more detail later). If the parameter is particularly impactful on the model's response, then fixing it to the apparent physical value implies an empirically worse representation of reality in your computer model. A prior distribution placed on such an input can and probably should be centered on the expected physical value since that is a defensible point of comparison. However, the final variance may be non-zero.

Residual variability is the discrepancy between the model prediction and value of the real process due to the fact that a real process will rarely arrive at exactly the same conclusion in repeated trials. This actually consists of two different types of uncertainty: the inherent unpredictability of the process, or variation that exists in due to a lack of control of the real process. Put another way, the latter is variation that could be reduced given more control over the process, but cannot be contained in the present circumstance. This is similar to model inadequacy, but in this case the value is averaged across these unrecognized conditions instead of being relative to the true mean of the process.

Parametric variability exists when we want to use the model to predict the process when some conditions are uncontrolled or unspecified. In this case, the inputs require more detail than we actually want to use or are capable of using. These input parameters are left unspecified and typically varied using some appropriate joint distribution, which adds uncertainty to the predictive process.

Observation error occurs in calibration because calibration is based on actual observations of a more credible source. Due to this, there must be a statement allowing for uncertainty in these observations. However, in reality, it is typically not feasible to separate

observation from residual variation.

Code uncertainty is related to practical implementation of codes. In a deterministic code, *in principle,* once an input vector is specified, the response is immediately designated. However, *in practice,* these models are typically quite complex. As such code execution takes some amount of time, from mere seconds to days or even weeks, to complete. Until such time as the code has converged for a particular instance, there is uncertainty in the response of that code. In the cases where model execution is particularly intensive, a full exploration is typically not practical so this uncertainty must be acknowledged. Methods of dealing with this uncertainty are related to the field of design of experiments and surrogate modeling (for more information see [47, 48, 49]). Designs of experiments are templates for efficient exploration of variable ranges to gather as much information in as few data points as possible. Combined with regression models one of the primary techniques is the screening design, wherein the model sensitivities are used to determine the critical parameters to be included in the exploration. Having a smaller set of variables is critical for thorough exploration of the parameter space. *Surrogate* models or *metamodels* are regressions that can be used to interpolate to points of the parameter ranges that have not yet been evaluated. The accuracy of the prediction of these regressions is dependent on the data currently available. Some regression techniques, such as *Gaussian process regression*, also called *Kriging*, and related techniques such as *Co-Kriging*, allow for regressions that also provide distributions instead of simple predictions[50, 51, 52, 53].

3.5.4 Types of Uncertainty: Riley and Grandhi

Building from the work of Kennedy and O'Hagan as well as others, Riley and Grandhi discuss model-induced uncertainties in a number of works. In these works, the sources of model-induced uncertainty are broken down into three different forms:

- Parametric uncertainty
- Predictive uncertainty

- Model-form or Model-selection uncertainty

These three different categories are unique but are closely coupled in their development and quantification. Due to this linking and the importance of each type of uncertainty, great care must be taken in determining which types to consider, how they are quantified, which are dominated by the others[54].

Parametric uncertainty in this context refers to the inherent uncertainty in the inputs and contains both aleatory and epistemic uncertainty[55]. There is some source for confusion here as in the Kennedy and O'Hagan description there is parameter uncertainty and parametric variability. Both of those instances refer to situations where the input values are uncertainty, either due to the need to calibrate the value to a specific instance, or due to a decision to vary an input over a joint distribution instead of assigning a single value for any given instance, respectively. In this case, the aim is uncertainty quantification instead of calibration, so parametric uncertainty is a term that jointly describe the inherent variation in defining the value of the input parameters.

Predictive uncertainty here denotes the discrepancy between model results and the described true physical scenario. The severity of this uncertainty is typically related to the models' ability to represent physical phenomena[55]. The process of validation is assessment of a computational simulation by comparison to experimental data[56]. As such, the predictive uncertainty is directly the type of uncertainty being addressed. When assessing predictive uncertainty it could be shown that the shape of the trend is correct but is simply shifted. This shows the need for calibration to move from predictive uncertainty assessment to validation. The existence of predictive uncertainty is a direct consequence of the simplifying assumptions that were made to represent the physical scenario mathematically and the effects of implementing such a model as a computer code. An example of such an assumption is an inviscid or incompressible flow assumption in aerodynamic analysis. Each of these would limit the applicability and predictive capability of the model.

Model-form or model-selection uncertainty comes up in a practical engineering context

when, as previously described, multiple models exist to represent the same system. For a particular scenario, it can be easily assumed that one of the models from the set best predicts the behavior of the system, a conclusion attributed to [57]. However, it is beyond the capability of the designer to have complete certainty as to whether a model provides the highest accuracy over the entire design space. Under the correct circumstances, it can be shown that a single model is more accurate than the others at a single point, but to prove that this quality can be extrapolated to the rest of the design space, all of the knowledge must be known about the entire design space. The problem here is that if full knowledge is known to justify model selection throughout the design space then the need for modeling has been removed. Additionally, the situation of having perfect knowledge of the entire design space provides enough difficulty to be considered not practically possible[55]. In certain circumstances where the behavior is comparatively simple, the "best" model may present itself throughout the process of verification, validation, and model use. However, most models are too complex to assume this will occur and, as discussed previously, benefits can be shown in the correct usage of the non-best models. This argument leads to the conclusion that while a single best model may theoretically exist and the comparative model-form uncertainties within the model set are of definite interest, setting the goal of the method as the selection of the single best overall model is an ill-advised path.

Observation *In a set of models there should exist a single model that is the most correct over the design space. However, the goal of quantification of model-form uncertainties should be to better understand the relative abilities of different models, not exclusively to downselect to a single universally "best" model.*

These descriptions of uncertainty aid in the discussion of model predictive capability and the relative comparisons in the process of model development and selection. The ability to specify the source of epistemic uncertainties aides in their reduction. Additionally, the categorization of aleatory uncertainty allows a better understanding of how to handle it in the uncertainty quantification process, as well as moving forward through verification,

validation, calibration, optimization, analysis, and design.

3.5.5 Uncertainty Quantification and Propagation

The quantification of uncertainties is a very active area of research. Table 3.2 shows some of the methods for propagation of uncertainty and how they have developed over time.

Table 3.2: Methods for Propagation of Uncertainty and Approximate Year of Appearance[36]

Name	Year
Wiener chaos expansion	1938
Monte Carlo method	1946
Quasi-Monte Carlo method	1961
Smolyak rule	1963
Latin Hyper Cube	1977
Generalized Polynomial Chaos	2003
Sparse Grids Quadratures	2003
Sparse Grid Pseudospectral approximations	2008

This work provides some background as well as descriptions of some of the most relevant methods for uncertainty quantification and propagation. It is not intended to be a complete review of all of the available techniques in the field of uncertainty analysis. A more general overview can be found in Bigoni's dissertation[36] or in the survey conducted by the Department of Energy[58].

While the types of uncertainties as defined by Riley and Grandhi are uniquely defined, they are not necessarily independent of each other. The nature of their unique qualities limits the wide application of certain methodologies to multiple types of uncertainties. Specifically in aeroelastic design, Riley and Grandhi list out some examples of the extensive work that has been done towards the quantification of parametric uncertainty, but the techniques used do not lend themselves towards the other types of model-induced uncertainty: predictive and model-form uncertainty[55].

The predominant issue in the undertaking of quantifying predictive uncertainty is that some information about the real scenario must be known. However, as described in the

category of observation error by Kennedy and O'Hagan, there are other issues presented when using experimental data points to represent the real scenario. These issues can be due to human or equipment errors as described in the endogenous uncertainties described by Robertson. Experimental data is also often developed using a physical model that is itself a "reduced-order" representation of the real physical system being described. The simplification of the physical model and the circumstances of the experiment such as the wind tunnel or other test stand being used add a layer of abstraction from the actual system that makes accurate quantification of predictive uncertainty that much more difficult.

In the calculation of model form uncertainty, the goal is typically to calculate a probability for each model in the set that it is the correct model. These probabilities must sum to one.

Bayesian Model Averaging

One of the techniques that has been put into use for the calculation of model uncertainties is Bayesian Model Averaging(BMA)[55]. The set of experimental data used is denoted by D, and the probability distribution for the adjusted model is represented as $Pr(y|D)$. This is an average of the posterior distributions of all of the models to be considered weighted by the probability that each model is the correct model. Put mathematically, this posterior distribution is described by 3.2 where y is the adjusted model of interest, and M_i is the i^{th} model.

$$Pr(y|D) = \sum_{i=1}^{N} Pr(M_i|D)Pr(y|M_i, D) \tag{3.2}$$

The calculation of the posterior model probability $Pr(M_i|D)$ for the current model uses Bayes Theory, as shown in 3.3.

$$Pr(M_i|D) = \frac{Pr(M_i) \times Pr(D|M_i)}{\sum_{j=1}^{N} Pr(M_j) \times Pr(D|M_j)} \tag{3.3}$$

The initial probability that the current model is the correct model $Pr(M_i)$ is simply based on the number of models being considered since no additional information has yet to be included. At the beginning, they are all given the same chance, to $Pr(M_i) = 1/N \ \forall i$. The marginal likelihood of the i^{th} model is then calculated using the experimental data as shown in Equation 3.4.

$$Pr(D|M_i) = \int Pr(D|\bar{x}, M_i) Pr(\bar{x}|M_i) d\bar{x} \tag{3.4}$$

In this equation, $Pr(\bar{x}|M_i)$ is the prior probability density of M_i and $Pr(D|\bar{x}, M_i)$ is the likelihood of M_i. It is noted that the integral in 3.4 can be difficult to compute in cases where the models have non-deterministic outputs, in which case something such as a Markov Chain Monte Carlo must be applied to approximate the integral instead. On the other hand, with a deterministic model, the posterior can be given as $Pr(y|M_i, D) = Normal(f_i(\bar{x}), \sigma_i^2)$, where σ_i^2 is the variance in the prediction function, $f_i(x)$, due to random error. This variance can be calculated by comparing model predictions to the experimental data points d_k, as shown in Equation 3.5.

$$\sigma_i^2 = \frac{\sum_{k=1}^{m} (d_k - f_i(x_k))}{m} \tag{3.5}$$

The posterior mean and variance of the adjusted model, shown as y, is shown in Equations 3.6 and 3.7.

$$\mathbb{E}(Pr(y|D)) = \sum_{i=1}^{N} Pr(M_i|D)\, \mathbb{E}(Pr(y|M_i, D)) \tag{3.6}$$

$$\begin{aligned} Var(Pr(y|D) = &\sum_{i=1}^{N} Pr(M_i|D) Var(Pr(y|M_i, D))^2 + \sum_{i=1}^{N} Pr(M_i|D) \\ &\times (\mathbb{E}(Pr(y|M_i, D)) - \mathbb{E}(Pr(y|D)))^2 \end{aligned} \tag{3.7}$$

One of the primary benefits of using Bayesian Model Averaging is that both predictive and model-form uncertainties are taken into account. While the Bayesian Model Averaging approach has proven useful for predicting the model probabilities in the quantification of model-form uncertainty, there are couple of issues. One of the limitations is that the model form uncertainty is simply represented here by single numbers in the model probability. This provides a straightforward approach to the comparison but there is more to the difference between the models in the set than can be represented by a single number of model probability. This will be discussed more later. Secondly, there is the reliance on experimental data represented by D to calculate the model probabilities. Some amount of verification and validation must be done on the models in the model set in order to justify their inclusion. However, as is especially the case with conceptual and preliminary design, experimental data relevant to the current design interest is quite often not available and obtaining such data would require additional work toward a concept that typically has not been proven to the point of justifying the expense. The adjustment factors approach is an attempt at handling the reliance on experimental data and will be discussed in more detail now.

Adjustment Factors Approach

The adjustment factors approach enables further understanding of the model probabilities by focusing more on the interrelationships between the models instead of a comparison to experimental data. In lieu of experimental data, the model probabilities are supplemented using expert elicitation. Bayes theorem is still employed to modify the model results to account for the model-induced uncertainty present in the set. The adjustment factors approach has been used in a number of instances, specifically related to risk and decision-making in the nuclear field[55]. There are some variations on the adjustment factors approach in the literature based on type of factor being applied. This factor adjusts the distribution of the model that was deemed as best by the experts. The type used by Riley and Grandhi is an

additive adjustment factor, where the value of the adjusted model y is represented by the deterministic output of the "best" model y^*, plus the expected value of the adjustment factor E_a^*. This relationship is shown in Equation 3.8. If the adjustment factor is assumed to be normally distributed, as is it often is in lieu of more information, and since the model result is deterministic, the adjustment model will be normally distributed as well. Therefore, the variance of the adjusted model is simply the variance of the adjustment factor. Other adjustment factor approaches, such as a lognormal approach, exist in the literature.

$$\mathbb{E}(y) = y^* + \mathbb{E}(E_a^*) \tag{3.8}$$

Since the adjustment factors approach involves replacing experimental data with expert opinion, there is more uncertainty present due to the responses of the disciplinary authority. The modified adjustment factors approach was developed to attempt to account for this. Instead of the expert provided model probabilities being treated as deterministic values, they are instead treated as normally distributed random variables, as shown in Equation 3.9.

$$P(M_i) = N(P(M_i)_{exp}, \sigma_i), \text{ where } \sigma_i = min[0.05, 0.25 \times P(M_i)] \tag{3.9}$$

Essentially, the model probabilities are varied around the values provided by the experts using Monte Carlo sampling to obtain a set of model probabilities. These sets of model probabilities are used in the traditional adjustment factors approach to provide a set of adjusted models. In that instance, a resulting set of distributions was gathered using a Metropolis Chain implementation of a Markov Chain Monte Carlo sampling approach. The result is referred to as the aggregate adjusted model, y_{maf}, which can be compared to y, the original adjustment factors adjusted model. The difference between the modified distribution and the original, expert driven distribution are compared using the Bhattacharyya distance, a measure of geometric similarity between two distributions, which is described in Equation 3.10. If the Bhattacharyya distance is one, the two models are identically

distributed.

$$BC\left(f_{x_1}, f_{x_2}\right) = \int_{-\infty}^{\infty} \sqrt{f_{x_1}(x) f_{x_2}(x)} dx \tag{3.10}$$

The information gleaned from variation from the expert-provided model probabilities in the modified adjustment factors approach, however, is not a new set of model probabilities. It is an assessment of the sensitivity of the result to the model probabilities that are provided. Because of this, the adjustment factors and modified adjustment factors approaches are presented as the first step in the process of model-form and predictive uncertainty quantification. This is beneficial in that in the absence of experimental data, as is typical in early design, qualitative assessment is used to denote the ranking of model quality. This is acceptable because it could be reasoned that unless the models are incredibly complicated an expert should be able to put them in the correct order from worst to best. However, the determination of "by how much" is a different proposition.

The modified adjustment factors approach then tells you that if the Bhattacharyya distance does not meet some critical tolerance that has been set, then the model ranking is sensitive to the provided model probabilities. The downside is that the only determination that can be made from a poor variance between the two adjusted models is that more data is needed to form an accurate assessment of model probabilities. As stated before, the benefit of incorporating experimental data and using Bayesian Model Averaging is the ability to also quantify the predictive uncertainty. However, to reiterate, data is still being required that may not yet exist or is cost prohibitive to generate. If the Bhattacharyya distance is accepted, however, the adjusted model is insensitive to the given model probabilities and they can be accepted moving forward.

Unfortunately for this work, the primary benefit of this is not in the calculation of model-form error, but in the reduction of necessity of experimental data points. While this is an important consideration, it is only of a certain benefit in describing the difference in quality between models to be used early in the design process. What can be considered, however, is how the qualitative comparison between models by experts can be improved.

Quality of expertly elicited assessments can be improved through formal procedures, where protocols are explicitly defined and problems are decomposed to the point where a believable assessment can be made[59]. In this case, the comparison of physics-based models for engineering design and analysis is related to the description of model fidelity, which will be discussed in more detail in Chapter 4.

3.5.6 Model Understanding Through Sensitivity Analysis

The methods involved in sensitivity analysis pervade the processes of model development and general understanding. Put one way, "sensitivity analysis is required for understanding the extent to which a model is complicated enough but not too complicated[28]." Sensitivity analysis is relevant any time there is a model involved: scientific modeling, decision-support, financial and economic prediction, and any other system simulation[60]. Rabitz[61] stated that "the judicious application of sensitivity analysis techniques appears to be the key ingredient to draw out the maximum capabilities of mathematical modeling." Ferretti et al. specifically published an assessment in 2016 showing that in the academic scientific literature, use of sensitivity analysis techniques has risen notable in the past decade. While the increased use of sensitivity analysis is promising, much of this work still made use of simple methods that leave a lot to be desired. It should be noted that this assessment did not include engineering literature[62]. Borgonovo and Plischke provide a survey of the literature of the use of sensitivity analysis[60]. In that work they point to a number of previous reviews that were performed between 1997 and 2013. A general reference for uncertainty analysis can be found in the textbooks of Saltelli et al.[63].

The increase in computing power has led to a beneficial parallel increase in the complexity of models and in the ability to understand those models using sensitivity analysis. Code uncertainty as described by Kennedy and O'Hagan states that while the response of a model may be deterministic, model complexity prohibits reduced output uncertainty prior to executing the code for a given set of inputs. The complexity of models additionally ob-

scures the source of output variation with respect to model inputs. Put another way, it is beyond the intuition of even an expert to know a priori which parameters will contribute most to the variation of the outputs. Related to model-form uncertainty quantification and propagation, sensitivity analysis has been proposed as a potential guide to model fidelity improvement among other benefits[64].

There are different types of sensitivity methods:

- *Quantitative and model free[65]* v. Tailored to a specific mathematical approach or problem
- Local v. Global

Model free methods are those that can be applied generally to a variety of models as they are not tied to the assumptions of the model, such as linearity. Tailored approaches can only be applied to specific types of models, such as linear programs. Tailored approaches should have a tendency toward efficiency since the characteristics of the model have been directly taken into account, but to allow for a general model comparison approach, the sensitivity technique must be model free. The concept of *local* versus *global* is somewhat less intuitive.

Local sensitivity analysis methods are carried out in a deterministic fashion, examining the behavior of the model around a point of interest. In this way, there are no probability distributions being applied to the model inputs, and the model inputs are varied around a specific point of the design space. These local sensitivities can be elicited in very straightforward ways, such as "one-at-a-time" sensitivity analysis. This is the method where the parameters of interest are perturbed individually and their impacts on the model responses. The number of predictions that must be made scales linearly with the number of inputs, but since only one input is being changed at a time any interactions between variables are ignored. As model complexity grows, this becomes increasingly insufficient as interactions play a large role in the variability of the model. To counteract this other methods have been

developed, such as scenario decomposition, one way sensitivity functions, and differentiation based methods. Sensitivity analyses that are deterministic can seem somewhat limited, but their usefulness has been argued, especially in the absence of confident assignment of distributions to variables. They also tend to be simpler to implement and undertake, as they often require fewer model predictions.

While the limitations of one-at-a-time analysis is well-documented, it has been shown that in the bulk of scientific literature, it is the predominant method used. One of the main downsides relates to what is called the curse of dimensionality. As the number of parameters increases, the "mass of a hyper-cube tends to concentrate in its edges and corners." By forgoing inclusion of the majority of the design space, it is difficult to justify that the behavior of a model has been captured. Additonally, not only does evaluating sensitivity in a "one-at-a-time" way leave all of the interactions dormant, aforementioned theories such as design of experiments are designed to take these into account. As such, the fact that a large portion of scientific literature does not account at all for interactions is surprising[62].

Global sensitivity analysis, on the other hand, assess the sensitivity of the model via a probabilistic assessment. Probability distributions, either joint or marginal, with or without correlation, are assigned to the model inputs. The literature contains a rich set of methods for global sensitivity analysis, many of which are covered by Borgonovo and Plische: regression based methods, variance-based methods, density-based methods, transformation invariant methods, value of information based methods, and Monte Carlo filtering techniques[60]. The field of global sensitivity analysis is a robust and flourishing one, and since they take into account the entire range of variation and work with the related field of uncertainty quantification, it should continue to grow in its maturity and usage[66].

Screening methods are an important category of sensitivity analysis techniques and are designed to evaluate sensitivities while exploring the applicable ranges of the inputs. Since the entire design space is taken into account, some of the techniques of designs of experiments, mentioned earlier in this work, are used to maintain a "parsimonious number of

model evaluations[60]." Screening designs are crucial to model development, selection, and implementation. Complex models typically have a large number of parameters that can be varied. However, the methods used to understand models and explore design spaces are often heavily dependent on the number of parameters to be varied. Another term for screening is parameter reduction, which allows for the assessment of parameters to determine which are not statistically significant to the variability of the responses. As such, if domain experts agree with the statistical assessment, then the parameters can be "eliminated" by setting to a nominal value[28].

One set of commonly used methods are based on the process of Analysis of Variance (*ANOVA*). As the name would suggest, ANOVA and the methods based on it are statistical characteristics for assessing variance and variability of model characteristics. These techniques apply to many different applications of sensivitity analysis, including screening or variable reduction techniques. Functional ANOVA is "at the basis of the high dimensional model representation theory, which plays a fundamental role in global sensitivity analysis[60]."

Chapter 2 aimed to define models and the associated terms and categories to be applied in this context to mathematical models. Specifically, models implemented as computer codes for the purpose of design and analysis of engineered or engineering systems. This chapter has worked to extend the discussion of models and codes to the discussion of credibility assessment. The standard processes for credibility assessment have been defined: verification, validation, calibration, and accreditation. An overview of the types and sources of uncertainty as well as references for techniques to quantify uncertainties has been provided. Limitations to many of the standard techniques for model credibility with respect to the models and codes of interest are discussed. These are the lack of data typical in early design, design of revolutionary concepts, or incorporation of new techniques. Model understanding through sensitivity analysis enables improved model understanding due to comparison to the model or models themselves instead of some validated real dataset.

It is important to note that surrogate modeling regression, or model emulation techniques are crucial to many of these quantitative evaluations[67]. Probabilistic techniques for sensitivity analysis and uncertainty quantification often require many model prediction points. Add to that the number of evaluations required for the current application (e.g. optimization or calibration) with the need for repeated evaluations to achieve multidisciplinary convergence, and the computational requirements can be very large. This is even the case for codes that run in a seemingly reasonable amount of time on their own. As such, surrogates allow for an analytical predictor to be developed that can, with enough training points, stand in for the actual code, and be evaluated nearly instantaneously.

Similarly to this work, Uusitalo et al. present many of the same categories in the assessment of uncertainty for decision support in environmental modeling: expert assessment, model sensitivity analysis, model emulation, temporal or spatial variability, use of multiple models, and data-based approaches[67]. The specific methods used within each of those categories must be reevaluated for application to a different area of expertise, specifically aerospace structural design, but the overall method is essentially the same.

3.6 Conclusion and Overarching Hypothesis

Proving the credibility of a model prediction essentially boils down to proving to an expert that the generated results are acceptable. If an abundance of model and validation data is available, an appropriate quantitative method can do a great deal to assuage disbelief and instill confidence. However, if high quality validation data were generally available, modeling would not be necessary, and the problems discussed in the preceding chapters could be easily avoided. Given the difficulty of developing a trusted model of any complexity, model data in a large multifidelity set is far from a guarantee.

Despite a lack of data, models are still commonly selected based on availability, which can lead to type I-III errors. To avoid this as much as possible, a detailed investigation of the qualities of models that affects its fidelity is needed to improve the way the probability

of being the highest fidelity is calculated in a data-independent way. When model data is available, it should also be leveraged, even in the absence of validation data. Following that, the understanding of relative fidelity should be combined with even a rough approximation of model cost to allow for modeling decisions to be informed by both fidelity and efficiency. This leads to the overarching hypothesis of this work.

Overarching Hypothesis *If a hybrid approach combining model fidelity heuristics with quantitative data comparison techniques is used to assess model fidelity based on expert opinion and available model data, more informed model selections can be made for use throughout the process of design.*

The first aspect that needs to be addressed is the understanding of model fidelity. Chapter 4 details an investigation into the definition and description of model fidelity in the literature to develop a new fidelity framework that provides clearer, more intuitive, heuristics for ranking the relative quality of models.

CHAPTER 4
DESCRIPTION OF FIDELITY

4.1 Research Question 1

As mentioned in the previous chapters, the process of ranking or scoring models in a set relates to the assessment of model form or model selection uncertainty. This entails the generation of model probabilities: typically via expert opinion or comparison to experimental data. Also, when thoroughly investigating a design space, trustworthy data is scarce, so there is a need to rely more on expert opinion methods of model assessment. In model form uncertainty, this probability is described as the probability that a particular model is the most likely in the set to be capable of generating the validation data. Put more generally, it could be described as a representation of the probability that a given model is the highest fidelity in the set. All of this leads to the development of the first research question of this work. When consulting an authority, the requirements for model form methods are restrictive and not particularly intuitive. They are asked to provide specific values for each model's probability, meaning they are directly assessing the order and magnitude of the difference between the fidelity of each model. Regardless of the way these opinions are combined, the basic method of inquest leaves a lot to be desired. This leads to the Research Question 1.1.

Research Question 1 *How can the process of generating model fidelity assessments early in the course of design be improved?*

It should be noted that the inclusion of a model in a set of possible models could imply that a model has already been developed, is being developed, or is simple enough that a plan for development is easy to produce. For this work, the assumption is being made that a model is at least in the process of being verified, and as such, some amount of data has been

generated or can be generated. This can be a difficult assumption to make, given that model development can be a costly and time-consuming process, but that will be discussed more in a later chapter. The most important aspect of note is that the methodologies developed in this work intend to take the most advantage of whatever information is available at a given point, and be flexible enough to allow for updates as more data becomes is generated. Given this assumption and the interest in early design, the most viable areas for improvement in model fidelity assessment can be subdivided into two categories:

1. How can the question "which of these models has the highest fidelity" be made more clear?

2. How can the process of fidelity ranking be made more intuitive for the experts, and flexible enough to take advantage of available model data?

The first of these questions is the focus of the rest of this chapter.

4.2 Introduction to Fidelity

The description of fidelity is related to the description of modeling and simulation, but with a notable exception. Describing models, as covered in Chapter 2, pertains to the aspects of the model, how it is developed, how data is transferred, how the equations are solved, etc. Model credibility, as described in Chapter 3, predominantly refers to the quantitative methods of exercising a model to make sure the requirements have been met and examine the results. The description of fidelity combines these two topics. The accuracy of the representation of reality in the model is intrinsically dependent on the model type, but the description itself is irrespective of the model's type. Model fidelity description and estimation is a way of reporting how well the system represents reality in order to predict how well reality will be represented by the model.

While it is important to define the fidelity of a model the driving purpose in model selection is also important. It is commonly known that "most engineering systems can be

approximated with models of varying degrees of accuracy or *fidelity*." However, the naive opinion stated in the context of model selection is that "all things being equal, it would be desirable to use the most accurate model[15]." The difficulty with this statement is that it is in direct contrast with Occam's Razor, which roughly states that the best solution is the simplest one that is valid. Many authors recognize that there is such a thing as too much fidelity[68]. There is benefit in using models at various levels of fidelity alone or in conjunction throughout the design process, so enabling the comparison the fidelity of these models is important.

The real reason why using the absolute most accurate model at all times is not desirable is that the accuracy of the model is only important to the aspect or aspects of interest to the current application or need[68, 2]. For example, if designing an aircraft with a turbofan engine and the only required information is required thrust, a low fidelity model may be sufficient to identify the correct attributes. However, if the characteristics of the individual subcomponents are needed, then a high fidelity model would enable the engine to be parameterized by internal geometry and materials[69]. Thunnissen provides another example, in the case of designing three different scale aircraft models: one for a three-year old, one for a ten-year old, and one for wind-tunnel testing. For a small child the most important parameter may not be accuracy at all, but to make something "fun." This soft requirement could be translated as simplicity and durability. In the case of the older child's toy, the requirement could be that the model is "representative," simply meaning visually similar in scale. For the wind tunnel test model, the parameter of interest may be "exactitude," which is similar to representative but requires an additional level of detail to the exact in-flight shape and surface finish appropriate for physical testing. There is no point in requiring a more complex model than is necessary given the need. One must be especially prudent in selection, as it has been noted that the computational cost for a high-fidelity model could easily exceed 100 times that of a low-fidelity model[2].

Ideally the uncertainties that a designer would face would be related to the parameters,

but as discussed previously, a non-trivial amount of that uncertainty is due to other sources. An example is in mass estimation where the system value may be dependent on component availability, material properties, requirements, but also the fidelity of the available models[2].

4.3 Definition of Fidelity

Many definitions for the term fidelity in the context of modeling and simulation are provided throughout the literature, a sampling of which are shown here:

[70] The degree to which the model produces the same outcomes as the tangible, physical system

[68] "People tend to use the term fidelity as a kind of shorthand for describing how closely a simulation corresponds to the 'real thing'"

[71] *(DoD M&S Glossary)* The accuracy of the representation when compared to the real world

[72]
1. The degree to which a model or simulation reproduces the state an behavior of a real world object or the perception of a real world object, feature, condition, or chosen standard in a perceivable manner; a measure of the realism of a model or simulation. Fidelity should generally be described with respect to the measures, standards, perceptions use in assessing or stating it.
2. The methods, metrics, and descriptions of models or simulations used to compare those models or simulations to their real world references or to other simulations in such terms as accuracy, scope, resolution, level of detail, level of abstraction and repeatability. Fidelity can characterize the representations of a model, a simulation, the data used by a simulation (e.g. input, characteristics or parametric), or an exercise. Each of these fidelity types has different implications for the applications that employ these representations.

Most of these definitions refer to the fact that model fidelity is related to the question of validity in that the fidelity of the model should represent how well reality is being represented. This begs the question: If the majority of the definitions for the term fidelity agree with each other, then why is it such a difficult thing to describe?

4.4 Difficulty in Defining and Describing Fidelity

Ambiguity Due to Ubiquitous Usage

One of the primary difficulties in defining and describing fidelity is the term's ubiquitous use in a variety of contexts. Many works simply use fidelity as a surrogate for accuracy, which is not out of line with most of the definitions in the context of modeling (e.g. [73, 35]). In other instances, the term is used as a substitute for the level of detail present in a model (e.g. [74, 2]). Numerous works use the term fidelity simply to refer to "low-fidelity" and "high-fidelity" models, which implies the models are different in some way but doesn't define how they are different. This is a vague usage and will be discussed more in Section 4.4.1.

Trucano et al. describe verification as "mathematical accuracy" but validation as "physical fidelity." This is similar to the representation of fidelity as equivalent to truth and accuracy, and they make note of the lack of scientific fidelity and inaccuracies that can appear. Yet another way that the term fidelity frequently comes up is in reference to the fidelity of a regression equation. In a case such as [75] it would typically appear that the fidelity of a regression should simply refer to its predictive capability. However, fidelity could also be used to refer to the flexibility of a regression technique, similar to the model inadequacy uncertainty discussed in Section 3.5.

Despite the fact that much effort has been into defining fidelity and its relevant terms, almost every major work still describes the state of the art as being ill-defined. In the context of Distributed Interactive Simulation, Lane and Alluisi described the difficulty of defining fidelity. The increase in literature actually confused the issue, presenting more than

twenty-two different types of fidelity (e.g. physical, equipment, psychological, perceptual, functional, procedural, task, logistic, threat, etc.)[68]. The widespread usage of the term makes it difficult to even determine the correct literature to reference. In one case, fidelity an resolution can refer to the accuracy of computer generated images for the purpose of animation or games. This is different because plausibility is important instead of accuracy, though if either resolution or behavior is too far off, it can be a detriment to the viewer[76]. On the other hand, there are "authorities [who] think that the term, fidelity, should apply only to *hardware* (does it look like and operate like the actual equipment)[68]."

Referent Difficulty

Another predominant difficulty in defining fidelity is related to the discussion of model credibility in Chapter 3. Accuracy or ability to predict reality is related to the definition of a *reference situation*[68]. If observations can be generated, then there is the potential for observation error, whether in the undertaking of exercising the system, the measurement of the system, or at some other point in the process. There must also be a process of validation at each point in the process to properly compare the behavior instead of simply the pure number produced by the observation. This is related to the model variability and is part of the purpose of calibration, since the value in the un-calibrated model may not directly relate to that value in the physical system. Exacerbating that issue is a myriad of reasons why it may not even be possible to obtain an observation in the first place, some of which are as follows (adapted from [70]):

- The physical system does not yet exist, a problem that is quite common in design, especially in the case of radically new designs or designs that incorporate new technologies
- It is hazardous to the viewer the to obtain observation data of the system while in operation

- It is hazardous to the system to obtain observation data while in operation
- Organizational or ownership issues come into play to prevent observation of the system
- Observation will negatively affect the system behavior, similar to the Heisenberg principle
- Operating environment does not yet exist
- Operating environment is too "dirty" or inconsistent for reliable measurement
- Models are "steady-state" which is not attainable by most physical systems

The case of the environment not yet existing is one that is not typically mentioned, but is especially important to this work. The environment that an aerospace vehicle operates in is quite commonly a complex and difficult to attain set of circumstances. The two primary contributing factors to this are altitude and speed, which create environments under which measurement is not easily attained. Additionally, aleatory day-of-flight uncertainties exist such as exact atmospheric conditions which increase the complexity. The difficulty of observing an aerospace vehicle in flight is one of the reasons for wind-tunnel testing, but this creates additional complications such as the use of physical models instead of the real system.

When discussing model-form uncertainty in a quantitative sense, fidelity is essentially defined in relation to the model probabilities described in Section 3.5.5. However, model-form uncertainty is related to the complications of defining fidelity as the uncertainties are often difficult to define. Some example of some of these complications include different model processors, theories, assumptions, mesh sizes, or boundary conditions[77].

4.4.1 Fidelity as a Scale

In numerous works fidelity is simply used in an ad hoc manner to describe whether a model is low-fidelity or high-fidelity. This is sometimes used in reference to comparing specific examples, e.g. high-fidelity referring to a detailed representation of a single bolt where low-fidelity refers to a generic description of the entire structure[78]. This loose scale is somtimes used to compare categories of models or software, and sometimes without providing any example for the comparison.

Generally using the terms low-fidelity and high-fidelity to refer to two competing models is not automatically a misuse of the term. The models are assumed to be verified and in certain circumstances the differences between them can be clearly designated based on previous experience. It is implied that the "high-fidelity" model describes the system in more detail, with a more accurate representation of the physical behavior, and the model has an increased computational expense. If all of this is the case then the model may do a better job of predicting the behavior over the whole design space. However, any actual analysis to prove that this is the current circumstance occurred prior is typically left out. Due to this, the statement that one model is higher fidelity than the other is simply implied and for as much as the reader knows, was based entirely on a loose subjective comparison. As such the assessment of fidelity may not be incorrect but the lack of information on it undermines the credibility of the model selection.

There are inherent limitations added to the situation by applying a specific category or scale to fidelity. When one technique or piece of software is described as high-fidelity compared to another, there are a variety of subjective implications. If, for example, finite-element modeling is referred to as high-fidelity, then the person providing that description is implicitly referring to a specific type of model that they have developed in a finite-element framework. In reality, finite-element modeling could actually refer to any number of levels of fidelity. The model could be a one-dimensional element representation of a beam or beam-like structure such as a tower, launch vehicle, or high aspect ratio wing. It could also

refer to a two-dimensional shell model of a thin-walled structure, such as most aerospace vehicle structures: wings, fuselages, or launch vehicle tanks and barrels. The person providing the description could also be referring to a three-dimensional solid element detailed representation of a structure; this could include an engine block, a bolt, or most any other structural part. Also left out of this description is the type of solution being used with the finite-element model: material linearity, solution linearity, time-dependency, etc[79]. Most of the more detailed models listed here would be well out of the realm of someone providing weight estimation in the early phases of aircraft design, as described in the categories of weight estimation models provided by Roskam[80]. On the other hand, to those using finite-element modeling for the representation of material behavior, performing a linear static beam or even shell analysis is far too simple to ever spend much time considering.

Using "subjective adjectives" to describe fidelity is still a practice in fairly common usage in some parts of the literature. However, even in 2000 Roza et al. commented that in the available modeling and simulation literature, there is "community wide recognition" that using simple qualitative terms such as low, medium, and high "can no longer fulfill the current simulation requirements[81]." As well, the Simulation Interoperability Standards Organization's (SISO's) Fidelity Implementation Study Group (ISG) warned against such "single point or qualitative descriptions" such as a simple low, medium, or high to describe model fidelity in their seminal publications based on the Simulation Interoperability Workshop (SIW)[82, 71, 72].

Observation 1.1.1 *A fidelity description needs to be scale-independent to remove the subjectivity applied of an expert, implying more detail than can be put forth in a simple categorical representation. The goal should be to enable one-to-one comparisons based on the aspects of fidelity.*

4.4.2 Fidelity as a Standard

Much of the work that has been done towards the description of simulation fidelity has been with the intent of producing fidelity standards, or at least the framework for the definition of a fidelity standard within a given discipline. To do so, one of the primary things that has been provided is clear definitions of the terms used to describe fidelity. This is crucial to enable straight, to-the-point, conversations between those describing fidelity instead of enduring great confusion by discussing the same thing using slightly different terms.

Observation 1.1.2 *A description of fidelity must adhere to very specific linguistic usage. The terms used to describe fidelity and the sources of fidelity must be based on agreed-upon definitions to alleviate confusion between the different groups involved in model development, selection, and execution.*

Standards are needed in cases where consensus is needed but they should not restrain creativity in design. The definition of a standard for fidelity has been proposed predominantly in the simulation community for the development of training simulations and systems of systems simulation, so that a standard referent and simulation template could be developed. Specifications include items such as: physical characteristics, functional characteristics, intrinsic quality characteristics, and extrinsic quality characteristics. The benefit is easier identification, verification, and validation so that meaningful work can be done with less upfront cost. However, the development of a standard referent or template will always be somewhat discipline or application specific and requires the consensus of experts in that discipline or application field. As such this work is not intended to provide a standard, but rather aim to clarify the discussions that could lead to one.

4.4.3 Comments and Research Question 1.1

The difficulty in understanding fidelity creates a need for not only consistent definitions related to the terminology, but for a fidelity taxonomy. This is analogous to the description

of uncertainty referenced in Section 3.5 where a "sound taxonomy" is only categorized by its "fundamental essence," so that further discussion can be focused on sources of those types of, in that case, uncertainty[37, p. 51]. The extension of this discussion to model fidelity leads to the following research question:

Research Question 1.1 *What are the fundamental characteristics that drive model fidelity, analogous to the decomposition of uncertainty into aleatory and epistemic?*

4.5 Previous Fidelity Frameworks

The important aspects of the description of model fidelity are discussed in some works, especially those in the modeling and simulation literature. In order to determine the important characteristics it is useful to evaluate the model types and potential errors and uncertainties discussed in earlier chapters. As discussed in Section 2.2, one way of categorizing disciplinary design models is: regression of existing data, simple analytical approximations, analytical models, and discretized numerical methods. Some of the main trades in model usage and fidelity between these categories is the trade between flexibility and accuracy. This is related to the amount of detail present in the model, the assumptions that are made about the mathematical descriptions in the model, and the amount of the physical world that you are trying to capture. What follows is an assessment of some of the descriptions in the literature and the selection of key aspects for the description of fidelity to be put into practical usage.

Bailey and Kemple

[70] *Resolution* The degree to which detail is included in submodels

It has long been known that amount of detail is one of the aspects of the description of fidelity of a model. The topic of qualitative model credibility description has been called *conceptual or subjective validation* by Balci. Conceptual validation is sometimes thought

to be a bit of a contradictory term since one of the aims of validation is to remove the subjectivity of qualitative assessment. However, it has been repeatedly shown that the difficulties associated with design and validation create a need for a more flexible, qualitative credibility assessment. One of the primary goals of conceptual validation is to increase the resolution of models. To give an example of resolution increase in an object-oriented context[70]:

- Replacing larger objects with "semiautonomous" subobjects
- Replacing simple decision logic with more complex logic
- Using more source data (e.g. higher resolution terrain data)
- Including more objects
- Improving approximations

As the resolution is increased, there is a need for more data. Increased resolution corresponds to an increased number of items or an increased number of aspects of those items that must be specified. Increasing the amount of data increases the associated risks inherent to data sources.

While model *resolution* or level of detail is an important aspect of the description of model fidelity, it must not be used as a surrogate for fidelity. Bailey and Kemple make note of a "hidden assumption" that model fidelity must not decrease with model resolution. This means that if model resolution is increased without sufficient consideration, model fidelity can decrease. They decsribe simulation modeling as the representation of objects, environments, and the associated interactions, whether between objects, between objects and environments, or internally. As the resolution is increased, there are more facets for which to describe interactions (e.g. smaller objects have more environmental dependencies). Put another way, simply increasing resolution does not necessarily increase credibility. This

can be a source of misunderstanding between managers and analysts, as an increase in resolution provides more apparent granularity, and as such, the appearance of credibility. "The sponsor cannot fathom why the analyst insists on ignoring physical realities of the system modeled, while the analyst sees resolution as a source of obfuscation."[70] While Bailey and Kemple describe the resolution aspect of fidelity in great detail other aspects such as the method of accounting for interactions between the entities is not as developed. For that, other sources are investigated.

Lane and Alluisi

Another perspective on the decomposition of fidelity is put forth by Bailey to categorize fidelity into three different terms: fidelity, realism, and validity. *Fidelity* in their context is "strictly an engineering term that refers only to the physical correspondence of the simulator's hardware to that of the actual equipment being simulated." This can be interpreted in the general sense to refer to what was being described above as resolution or level of detail. From the user's perspective, does the emulated version of the system appear the same as the true system. *Realism* here is related to the fact that the simulators in question are being used for training. Do the "perceptions" and "subjective judgements" of the users appear close enough to how it would appear in the real systems. While their viewpoint refers to the user-in-the-loop, it could be extended to refer to the realism of the behavior of entities being modeled. *Validity* is, similarly to the typical definition, the "suitability of the simulation for a specific application." For the particular application, they are interested in whether a simulator can allow for sufficient proficiency in training. Validity in a more general sense would refer to whether the physics being modeled represent the actual environment well enough for the prediction to be trusted. The categorical decomposition here was developed for a specific usage where there are individuals being trained by simulators and so many of the important factors relate to that human-in-the-loop quality and aesthetic characteristic. As such, when generalizing the description for use in physics-based modeling for design

and analysis, the factors seem to find too much overlap to be generally applicable.[68]

One of the points worth noting that is developed by Lane and Kemple is the topic of *fidelity anchoring*. While the specifics again are too specific to be generally applicable to the context of this dissertation, the point of the method is of great interest. *Fidelity anchoring* is the process of determining what level of fidelity is appropriate without providing too much fidelity and doing more work than is needed. The main problem with general applicability is that in order to anchor the fidelity, one must be aware of what the simulation is intended to accomplish and the probably range of applications for which it will be used. They based this on three aspects:[68]

- *Effectiveness* : related to validity
- *User acceptance* : presenting the simulator in a way that will make people want to use it
- *Affordability* : taking into account the man-power, computational requirements, and fiscal budget

As before, the user acceptance aspect is directly related to the human-in-the-loop aspect of simulation training, but the idea is still important. As mentioned in Chapter 3, at the end of the day the acceptance of any model comes down to the apparent level of credibility as presented to a decision-maker. The important aspect of the concept of *fidelity anchoring* is the ability to decompose the problem and assess the maximum model fidelity required at the current point in the design process.

Observation 1.1.3 *The process of describing the models at hand and comparing their fidelities is not intended to rule out lesser models, as a multifidelity environment has been shown in the literature to be beneficial for computational efficiency. Instead, an intended purpose should be to cap the achievable level of fidelity to be used at the current point in the design process so that an answer is achieved without unnecessary rigor.*

SISO Fidelity ISG

As mentioned earlier, a seminal work on the understanding of fidelity was provided by the SISO Fidelity ISG[82, 71, 72]. One of the primary purposes of their is to provide a set of clear and consistent definitions of the terms related to fidelity in modeling and simulation. It is important to clearly define a set of terms so that communication can easily occur between developers, users, managers, analysts, etc. Many of the terms used in the process of describing fidelity are common words with multiple definitions, and even in technical fields there is overlap between modeling and simulation terms and other software disciplines. The myriad subtly different definitions further complicate an already confusing issue, so all of the terms should be treated rigorously as described in Observation 1.1.2.

Four different fidelity frameworks are covered by the Fidelity ISG. The first of those is called *Fitness for Purpose*, and is based on two categories: resolution and accuracy. *Resolution* here is defined as "the extent to which the simulation models each aspect of the real world," and *accuracy* is "the agreement between the performance of these models of each aspect and the real world performance." The predominant method for defining model fidelity lies in the decomposition of the problem and assessment of the impact of each aspect on the two aspects, resolution and accuracy. This assessment is made by experts using a five-part scale: none, minimal, significant, substantial, and critical. The difference between this and defining fidelity purely on a qualitative scale (e.g. high, medium, low) is that the definitions in this application are applied to the decomposed aspects instead of the entire vehicle. This is in agreement with the Bailey and Kemple assessment that resolution, or level of detail, is an important aspect to how the model will behave. Accuracy, given this specific definition (which is not in complete agreement with the rest of the Fidelity ISG glossary) seems to refer to the interactions and behavioral representation of the real world that has been mentioned, but using the term accuracy for it is, in my opinion, too generic of a term. This is based on that fact that in many cases accuracy is used almost interchangeably with fidelity, so stating that one of the primary aspects of fidelity is accuracy seems

redundant on the face of it.[72]

The second framework, known as *Cascading Accuracy*, refers to the layers of assessment that must occur to get from the top layer, *reality*, through the *conceptual model*, to the bottom layer, *simulation implementation or federation*. This framework is primary concerned with accuracy, and as such fidelity is defined as the inverse of the error between the model and reality, represented by the inverse of the error as described in equation 4.1.[72]

$$Fidelity_{\text{Nth Model}} = ||(E_{MF})_N||^{-1} \tag{4.1}$$

The third framework discussed in that work is referred to as *Sources of Uncertainty*. This framework is based on the usage of systems engineering tools to decompose the problem and assess linkages. The process to enable fidelity analysis is enabled by three "dimensions:" *enumeration* of the scope and depth of the dimensions and attributes, *identification* of measures of effectiveness, performance, merit, etc., and *specification* of the relationships between the involved entities. While the framework itself is fairly disjointed, many important points are put forth in the enumeration of modeling and simulation aspects and usage of systems engineering methods. [72]

Lastly, the fourth framework discussed is referred to as the *Fidelity Differentials Framework*. This method puts forth its own definition of fidelity as "the extent to which the model reproduces the referent, along one or more aspects of interest." This definition makes the aforementioned important distinction that the aspects of interest are what is most important to the actual fidelity, although the total fidelity should be related to all of the aspects of interest instead of just a subset. This framework describes three aspects of fidelity:

- Existence
- Attributes
- Behavior

This framework was developed predominantly with discrete event simulation in mind, specifically for military strategic decision making. *Existence*, therefore, is described to be determined by the level of *aggregation*. This refers to how the extent of reality that is contained within one model entity (e.g. is one entity an individual soldier or a platoon). This could be perceived as somewhat confusing as aggregation is, in many of the other contexts, thought of as the necessary accumulation of all of the necessary data, instead of a higher aggregation meaning a lower accuracy model as is the case here. The *attributes* are developed based on the *clarification* of reality. In this context, clarification refers to the removal of detail that is deemed unnecessary (e.g. a soldier's commissioning date is not included in the model). This again is a conflicting definition in that clarification of the model aspects could be understood elsewhere as increasing the detail, whereas here it is the removal of detail. The third aspect *behavior* is built on the process of *simplification*. This is a more straightforward linguistic representation as the simplification of behavior leads to something that is easier to model but is a less accurate representation of reality. Despite the issues posed with the specific terminology used, this description of fidelity does seem to capture much of the characteristics of model fidelity: the amount of reality being modeled, the detail being represented of that reality, and the representation of the behavior of that reality. This will be discussed in more detail later. The techniques of uncertainty quantification, sensitivity analysis, and regression are put forth for application to this problem, though these techniques were already elaborated in the previous chapter.[72]

Moon and Hong

Moon and Hong present a more recent (2013) perspective on the description of fidelity through "mathematical and logical arguments." One of the primary attributes mentioned is *abstraction*. Similar to what has been mentioned previously, they refer to the fact that modeling and simulation is in essence and abstracted perspective on the real world. "Without any loss of detail, modeling is impossible." This similar to previously mentioned topics re-

lated to fidelity, such as the description of behavior and interactions between entities being simplified to represent them in a mathematical context. The importance cannot be overstated, as "abstraction is *the one mechanism* that enables [the modeling and simulation] community to do its work."[83]

The other primary attribute of model fidelity as put forth by Moon and Hong is *resolution*. In some instances in the past, this term has been used interchangeably with abstraction, but there is an important distinction, as resolution refers to the "level of details in the model" that drives the implementation. Similar to Bailey and Kemple, this is the apparent likeness of the model to reality.

At the most basic level, an increase in resolution and a decrease in abstraction would in turn yield a higher fidelity. However, this is a simple high-level theoretical perspective on the subject. Moon and Hong provide a literature review of the preliminaries associated with abstraction, resolution, and fidelity in model information, some of which are explicitly referenced earlier in this dissertation. The literature review is presented primarily to point out that while some in the literature do understand and distinguish between resolution and abstraction, the M&S community still lacks clear, agreed upon, operationalized definitions of these terms to some degree. In the case of abstraction, the computer science discipline has an explicit definition which follows.

- Abstraction: An abstraction, written $f : \sum_1 \implies \sum_2$, is a pair of formal systems $\langle \sum_1, \sum_2 \rangle$ with language Δ_1 and Δ_2, respectively and an effective total function $f_\Delta : \Delta_1 \to \Delta_2$

In this description, $\sum_1$ is reality and $\sum_2$ is a simulation model, and the Δs are the languages used to describe those. In this way, the modeling f enables the abstracted translation from one to the other. This is not, however, as simple of a process in the M&S scenario.[83]

Instead, a representation of how resolution, abstraction, and fidelity interact within model information is provided. M denotes the information within a model, CS repre-

sents the "complete scenario," or in other terms, the information provided by a real world referent. U represents universal information, or all information of the real world. The fidelity of the model is denoted by the amount of overlap of the model information with the complete scenario. Mathematical proofs are developed in the reference that lead to the propositions shown in Equations 4.2 and 4.3.

$$M_{HA} >_A M_{LA} \implies M_{LA} >_F M_{HA} \tag{4.2}$$

$$M_{HR} >_R M_{LR} \implies M_{HR} >_F M_{LR} \tag{4.3}$$

What is implied by these equations is summarized as follows. An increase in resolution or a decrease in abstraction will denote an increase in model information. However, they are presenting the argument that an increase in resolution returns a superset of the model information, while a decrease in abstraction is not necessarily a superset. Because of this, a higher resolution should imply a higher fidelity, whereas a decrease in abstraction does not necessarily mean an increase in fidelity.

However, the Moon and Hong framework is developed, as many of the others, with simulation in mind, whether training simulation or system of system simulations. It is also an almost exclusively theoretical description that they themselves note in simply intended to act as a starting point in the development of fidelity frameworks. One of the main limitations mentioned is that their theoretic framework portrays fidelity level in terms of model information, yet models that use the same information can have completely different results. This is mentioned here to mention that while they have laid thorough groundwork regarding resolution and abstraction, there is a question of whether other aspects need to be included for thorough description of the essence of model fidelity. This will be discussed further in the following section.

4.6 Compiling Fidelity Frameworks

Compiling the various descriptions of the aspects of fidelity, three categories emerge, as shown in Table 4.1.

1. How detailed is the description of the entities?	2. How do the entities behave and interact?	3. Which entities are represented and to what extent?
Level of detail	Interactions	Inclusion
Resolution	Relationships	Boundaries
Clarity of attributes	Abstraction	Existence in model
Depth	Assumptions	Scope
Granularity	Behavior	
Precision	Simplification	
Agreegation of entities	Logic	

Table 4.1: Compilation of Fidelity Aspects

4.6.1 Resolution

The first primary aspect will be referred to moving forward as the **resolution** of the model. As described by Moon and Hong, this refers to the "level of detail" of the representation of a system. Put another way, resolution could be described as the level with which the model resembles the actual sytem. This is a term commonly used in image processing with an equivalent definition. In maps, resolution determines the size of the smallest feature. In terms of aerospace structures, it describes what is distinguishable:

$$\text{Vehicle} \rightarrow \text{Assembly} \rightarrow \text{Component} \rightarrow \text{Sub-Component} \rightarrow \text{Part}$$

For example, can you simply tell that you have an aircraft? Can you tell the wing apart from the fuselage? Are the individual ribs in the wing defined? At the highest level of fidelity you would be able to use your model to generate drawings of the individual parts and fasteners.

It is clearly important that the level of detail exists to represent the features of interest.

In finite element analysis of structural defects, if the defect is smaller than the mesh density, then the results may be irrelevant. If different beam cross-sections are being traded, yet the finite element only represents the cross-section of the element by area value, the results will have little meaning. However, as described in Section 2.2, it is not everything. Models that are distorted or dissimilar from the physical system can still provide insight[17]. The complicated interrelationship between resolution and abstraction is the focus of the work of Moon and Hong[83]. However, resolution is important to accreditation. This is related to what has been called "face validation," since a decision-maker is more likely to believe a model that looks like the system[84].

4.6.2 Abstraction

The second primary driver of model fidelity should henceforth be called **abstraction**. This is a somewhat less intuitive category, though, to reiterate from earlier, "abstraction is *the one mechanism* that enables [the modeling and simulation] community to do its work."[83] In a physics-based model this would be described as simplifying assumptions. Certain aspects are assumed to have a negligible impact in order to represent the model mathematically. It is important to note that while increased resolution denotes higher fidelity, reduction of abstraction is the goal.

An example of abstraction in structural analysis would beam the beam theory assumptions, e.g. Euler-Bernoulli or Timoshenko. You are assuming first that the dimensions of the structure are "beam-like", i.e. much larger in one dimension than the other two. Following that, Euler-Bernoulli beam theory makes certain limiting assumptions about the behavior of the cross-section of the beam. Under the circumstances that these assumptions are valid, a relatively simple set of equations can be used to understand a great deal about the structure. However, it does not take much complexity in geometry, material properties, or loading for these assupmtions to lose their efficacy. Similarly in aerodynamics, inviscid and incompressible flow assumptions are often made even with the understand that they do

not bely the practical nature of a real flow.

The aspect of abstraction can also be extended to models that are not based on physics. In those cases, such as the operations modeling of a manufacturing floor, abstraction refers to the description of logic. For example, in operations analysis of a vehicle system, it could be stated that there is a set probability of a part failing. Reducing the abstraction, the probability of part failure could be different for different parts, dependent on the age of the part, how or where it was manufactured, even the exact circumstances of its manufacture and use.

The viability of abstracting away the impact of certain system aspects is related to the Pareto principle[85]. It states that the majority of a system's behavior is caused by a small subset of the system's aspects. This is typically referred to in terms of eighty percent of the behavior to twenty percent of the causes, but those percentages are a generalization. In essence it is simply stating that a model developer is allowed to ignore certain aspects of the system because they do not have a significant impact on the estimation of the response of interest.

4.6.3 Scope

The one primary aspect put forth in this work that was not in the framework of Moon and Hong is what shall be called **scope**. It is a relatively simple concept but very often overlooked, as discussed in following paragraphs. Scope, in general terms, is the amount of the overall system that is included in the simulated representation. In thermodynamic terms this akin to the control volume. In terms of maps, scope determines whether the map describes a city, county, state, country, continent, planet, solar system, galaxy, or the known universe. One of the main differentiating factors between resolution is the resolution is more of a continuous scale, while scope is a boolean: a particular aspect is either included or it isn't.

A complicating factor for noting the importance of scope is its complicated interre-

lationship with resolution. Resolution, in this new fidelity framework, is specifically the detail of the description of items that are in-scope. Adding items that were not included previously seemingly adds more detail about the system, but the resolution pertains to how those new items are described, while the scope is inherently changed by their inclusion.

One of the reasons for a lack of interest in scope is that many of the people who have discussed scope in the literature have a fixed scope, often the full scope of the problem. This is the case for those developing simulators, as the entire system must be taken into account, or strategic decision-makers, where the area of interest can be explicitly defined using a map.

Scope is sometimes overlooked because it is often implicitly defined. For modelers, as resolution is increased and abstraction decreased, scope is typically decreased. The reasoning for this is easy to understand, as an increase in the level of detail and complexity of the equations that need to be solved drive up computational intensity. By necessity, the scope is decreased to problem feasible in terms of runtime. Additionally, increasing scope requires the definition of additional parts at a corresponding level of resolution and abstraction, and such a description may not yet be available for those entities.

For designers, scope is also changed without much explicit awareness. At the beginning of the process, all of the decisions regarding a system are available, so the entirity of the scope is considered. Once more high-level decisions are made, however, the decisions that need to be made become more focused, so the scope of interest is naturally decreased.

Even though scope is often set by necessity, it is nonetheless important. Interdependencies that are ignored between the included components and excluded segments add uncertainty that can be difficult to quantify. Put another way, the abstraction of the included entities can be held constant, but if their relationship with out-of-scope entities is unknown, some amount of understanding it lost. These complications are especially salient for aerospace applications due to their tightly coupled nature. One common exaple is the analysis of an aircraft wing. The thin-walled structure of a wing is both flexible and re-

sponsible for generating most of the lift required to keep the vehicle in the air. Structural analysis is dependent on loads, but the loads generated are dependent on the shape of the structure at a given instant. This tightly coupled problem is known as aeroelasticity and is a common area of research. The aeroelastic problem is difficult enough when the primarily load-bearing structure of a wing and lifting surfaces are within the scope, but there are number of other factors to consider. There may be other secondary structures that contribute to the stiffness of the wing. Fuel stores in the wing and the engine, if wing-mounted, will have a significant effect on the wing's behavior. Additionally, while the wing often provides most of the vehicle's lift, finding the appropriate conditions to calculate loads often requires some understanding of the tail, which is another flexible lift-generating structure.

Other examples could be given for launch vehicles. In sizing the primary structures of a rocket, the propellant must be within scope to some degree. Cryogenic propellants are typically pressurized, and this internal pressure can be used to provide stiffness to tank structures. Liquid propellant-containing structures are even sometimes sized such that they cannot withstanding their own weight without internal pressurization[86]. Additionally, solid-rocket propellants contribute to the overall stiffness of their stages, and this contributing stiffness changes as the propellant burns away. Similarly to aircraft, sizing a structure requires loads, and in this case, those loads come from trajectories. However, the trajectory is dependent on a description of the vehicle. If the mass of the vehicle changes enough, the trajectory would have to be recalculated. This can become a problem when a model that is designed to do a final sizing of a launch vehicle upper stage is used for design space exploration. A final sizing model might have fixed loads since, fortunately, dry structure mass is much smaller than propellant mass for a launch vehicle[87]. However, if the size of the tanks are changed significantly, then the dry and propellant masses will change, which should propagate to the lower stages of the vehicle. This could drastically change the mass of the vehicle, requiring an update of the trajectory. However, a model of only that upper stage with fixed loads gives no indication as to when this update needs to occur.

This literature review, discussion, and included examples lead to the following hypothesis to address Research Question 1.1:

Hypothesis 1.1 *The primary aspects that drive model fidelity are the resolution of the system description, abstraction of real-world behavior, and scope of included entities.*

This new fidelity framework of resolution, abstraction, and scope, will be further examined in the following chapters. By using these three more intuitive aspects to understand models instead of simply the difficult-to-define term of fidelity, additional information can be gleaned. As mentioned in the literature search, the importance of resolution and abstraction has already been put forth, especially in the work of Moon and Hong. However, scope is frequently aliased with the other two or ignored altogether, so more examples will be given as to why it should included as one of the fundamental characteristics that drive model fidelity. The following chapter describes how this fidelity framework is used to understand the relative fidelities of models and develop methods that lead to an informed model decision-making process.

CHAPTER 5
DEVELOPING A METHODOLOGY FOR FIDELITY AND EFFICIENCY ASSESSMENT

5.1 Introduction

Decision-making with regards to modeling choices is both a qualitative and a quantitative endeavor. As discussed in Chapter 3, the preferred method of proving that a model is the proper is via comparison to a sufficient amount of credible, real-world data, but, as stated previously, if all of the necessary data already existed, there would be no need for modeling. Therefore, there is some level of subjectivity as to whether or not a model should be believed.

For every design and analysis problem, models must be selected. Often they are simply selected based on familiarity or availability without addressing the potential inadequacies unless they arise at a later time. Often a more complex model than necessary is selected to include an effect that an expert knows exists, whether or not it is important to the current problem. This work aims to address the problems of initial model decision-making by leveraging all of the available information in an efficient and traceable manner.

First, given a problem, the possible models that provide that response should be enumerated. This may include models from early through detailed design if applicable, as the model selection requirements vary. Expert opinion must be leveraged to come up with this list and to initially determine which models contain the appropriate phenomenology for the problem at hand. Even models that are familiar or easily available should be checked as thoroughly as possible by those with relevant expertise to avoid Type III error, model accreditor's risk.

Once a list of potential modeling options is established, understanding of the model set

via expert elicitation should be used to avoid Type I errors, or model builder's error. Since the development of a new model can be a lengthy process, the probability of highest fidelity described in the previous chapter can help decision-makers to understand the relative quality of models in the set.

The fidelity framework developed in the previous chapter consisting of resolution, abstraction, and scope can be used to improve descriptive fidelity assessment, as the question is fundamentally changed. Instead of simply referencing fidelity, attributes that are easier to define and contribute to the quality of a model are addressed. This chapter addresses how the process of quantifying probability of highest fidelity must be changed based on the use of this fidelity framework instead of a single metric to provide an initial assessment of fidelity.

Following that, if models exist from previous projects, are easy to stand up, or are at some point in the verification and validation process, some amount of model data may be available. A method is developed to make the most use model responses as they become available. This attempts to address Type II errors, also called model user's risk, as using even minimal quantitative data can point out defenciencies that could easily be overlooked in a qualitative assessment. Additionally, it fills the gap between purely expert opinion derived model fidelity probabilities and those that require extensive model and experimental data, such as Bayesian Model Combination. As data becomes available, it should also be used to check appropriateness to the current problem, as discussed in a preceding chapter, though this is not the primary focus of this work.

Once the methods for assessing relative model fidelity are available, they are leveraged to inform the decision-making process. A scoring method is developed to use the probability of highest fidelity to rank ordered model combinations. Additionally, methods are developed to score models in terms of cost or efficiency so that a multi-attribute decision can be made. The allowable fidelity and cost at a given point can act as requirements to determine which model or models appear most favorable for the problem at hand. The

following chapter will use these methods as part of a decision-making framework applied to a more realistic trade study use case.

5.2 Descriptive Model Fidelity Assessment

5.2.1 Research Question 1.2

Using the fidelity framework in Hypothesis 1.1, experts can assess models in clearer terms than simply fidelity itself. Instead of addressing fidelity as a broad term, the specific resolution, abstraction, and scope of each model can be compared. Now the other problem mentioned in Section 4.1, is restated here as Research Question 1.2:

Research Question 1.2 *How can the process of model fidelity assessment be made more streamlined and intuitive for experts by using the newly defined fidelity framework?*

5.2.2 Requirements and Hypothesis

Model development is a laborious process, and multifidelity model development even more intensive. Therefore, a method should seek to understand as much as early as possible about the available options. Individuals with relevant expertise need to be consulted regarding the possible options for modeling a system. At that time, they should also be consulted regarding the relative fidelities of the enumerated possibilities. As such, the process of gathering these opinions should not be unnecessarily difficult.

Requirement 1.2.1 *The process of capturing fidelity descriptions should not be unnecessarily difficult as to allow for assessment during the process of model definition.*

Information gleaned from expert consultation is meaningful since it is based on prior experience and understanding of the assumptions inherent to both the model and the system. However, it is still based on relative generalities, so the understanding of the model set can be further improved as data becomes available. As such, a methodology for fidelity

assessment should be flexible enough to incorporate fidelity assessment based on model estimates as they become available.

Requirement 1.2.2 *A methodology for predicting model fidelity rankings should be flexible enough to incorporate scorings from multiple opinions and sources.*

Additionally, due to the nature of expert elicitation, varying opinions may arise. Different people may believe different things based on their past experience or one expert may think that multiple options are likely to be the case. Due to this, the methodology should allow for multiple opinions on the same aspect. The methodology should even potentially allow for a relative weighting based on the confidence given to each opinion if that is available. If the ability to apply weightings exists, but there is no opinion as to the relative confidences, each opinion can be weighted equally by default.

Requirement 1.2.3 *The method of compiling fidelity evaluations should have the capability to weigh the confidence in the samples relative to each other.*

These requirements lead to the following hypothesis:

Hypothesis 1.2 *If model fidelity assessment via expert elicitation is streamlined and utilizes the understanding of fidelity through resolution, abstraction, and scope, the resulting model fidelity probabilities are generated in a more traceable manner.*

5.2.3 Model Ordering by Fidelity Attribute

Redefining fidelity into three categories that are more linguistically specific helps in the process of making expert elicitation more intuitive, but there is another step. As mentioned before, typically an expert is asked to provide a scores: proportional rankings of each model as a percentage of the whole. This requires unnecessary effort on the part of the expert to keep track of the running total, and becomes much more difficult as the size of the model set is increased.

There are other ways to elicit this information, but even when fidelity is defined as resolution, abstraction, and scope, the model as a whole is still being described in broad

terms. Due to this, it is difficult to state the exact proportional relationship between two models with much confidence. What can be said with greater confidence in simply that one model should rank higher, lower, or the same as another. It could be argued that, for resolution and scope, a proportional scale could be developed in certain circumstances. The same cannot generally be said for abstraction, as changes in the assumptions, logic, or physics cause non-linear changes in predicted behavior. However, even when a scale could be developed, this requires additional effort on the part of the assessor and naively implies that the fidelity aspect and the actual fidelity of the model have a linear relationship.

All of this leads to the statement that the resolution, abstraction, and scope of models in a set should simply be judged via a simple ranking method. For each category or type of fidelity, the models should be laid out in order from worst to best, keeping in mind that two models can be scored as equivalent. Based on a particular order, scores can be generated evenly between zero and one, described as in Algorithm 1. Sample scores given a notional set of four models are shown in Table 5.1. Each of these scores represent the understanding of a model, which includes some uncertainty. This leads to Requirement 1.2.4.

Requirement 1.2.4 *The assessment of fidelity represents a combination of predictions with some uncertainty and should therefore be represented as a distribution.*

This method is based on the lesser hypothesis that at this point in the process, an ordering is the most justifiable method of determining fidelity that an expert can provide. In the current methods, experts are required to define not only the order, but the specific ratio of fidelity betw een models. Since the scores are based on high-level metrics and defined as a single number for the entirity of the models' behavior, it is practically impossible to justify why one model is stated to be 2.0 times higher fidelity than another, as opposed to 2.1 or 1.9. This leads to Assertion 1.2.

Assertion 1.2 *At the level of expert-provided qualitative assessment, with respect to fundamental characteristics of fidelity, requiring the exact ratio between two models is beyond the level of what can confidently be stated. Experts can help to define the relative placement of models in the set, but scaling of magnitudes should be based on comparison of quantitative results once they become available.*

Algorithm 1 Generating Normalized Scores From Order

Require: List of grouped models, i.e. [[1], [2, 3], [4]]
  $i \leftarrow 1$
  for all group in order **do**
    for all model in group **do**
      $model\_score \leftarrow i$
    end for
    $i \leftarrow i + 1$
  end for
  $total \leftarrow sum(scores)$
  for all model in models **do**
    $model\_score \leftarrow model\_score/total$
  end for

Table 5.1: Calculated Scores Given Order

Description	Order	Scoring
Simple order	$[1], [2], [3], [4]$	0.1, 0.2, 0.3, 0.4
Model 2 and 3 similar	$[1], [2, 3], [4]$	$\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{3}{8}$
All equivalent	$[1, 2, 3, 4]$	$\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$

5.2.4 Fidelity Density Estimation

While fidelity assessment based on expert opinion via model ordering streamlines the process, the scores that are generated are now based on three aspects instead of one. Instead of directly undesrtanding the probability that a given model has the highest fidelity in the set, the experts are describing whether each model has the highest resolution, lowest abstraction, or highest scope in the set. Due to this, the combined description of fidelity is

some combination of the three aspects of fidelity detailed in Chapter 4. Even in the unlikely case that a single opinion is given for each aspect and the resulting scores perfectly agree, there is still some uncertainty inherent in the expert assessment process. As such, Requirement 1.2.4is derived. The generation of such a distribution based on a discrete sampling of scores leads to a type of probability density estimation method, specifically Kernel Density Estimation (KDE).

Kernel Density Estimation

KDE, also known as a Parzen-Rosenblatt window in some discplines[88, 89], is a method for estimating a probability density for a particular result based on a number of discrete samples. A type of base distribution, called a kernel, is selected. Since each fidelity score represents an estimate with some amount of trailing uncertainty, and due to its ubiquitous nature, a Gaussian or normal kernel is commonly used, and will be used for this work. The Gaussian kernel is described in Equation 5.1, where x is a sample value and h is the bandwidth. A symmetric distribution of the selected type is centered on each sample and the overall density estimate is generated as the sum of the individual distributions, where the density estimate at a point y is described in Equation 5.2. An example of these sample disitributions and the resulting kernel density estimate are shown in Figure 5.1. These equations and more information on the KDE implementation used for this work is from the SciKit-Learn API and documentation [90, 91].

$$K(x; h) \propto exp\left(-\frac{x^2}{2h^2}\right) \tag{5.1}$$

$$\rho_K(y) = \sum_{i=1}^{N} K((y - x_i)/h) \tag{5.2}$$

The variance of the individual distributions is referred to in KDE as the bandwidth. The process of determining an appropriate bandwidth given a set of samples is an active

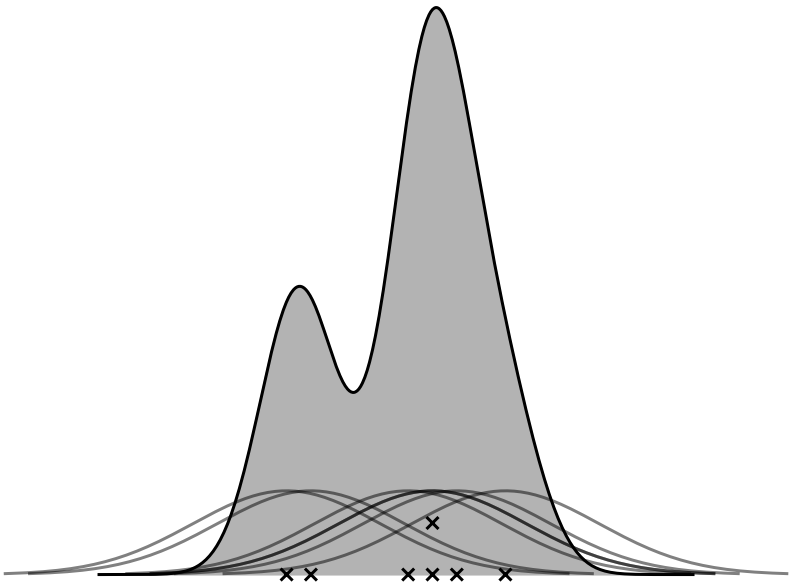


Figure 5.1: Kernel Density Estimation: Kernel Distributions at Samples That Sum to Density Estimate

area of research. If the user has an understanding of how the resulting distribution should appear, the bandwidth can be explicitly specified. Other simple method, such as rules of thumb, are often used. For this work, a more generalized numerical approach is taken, generally referred to as cross-validation. Cross-validation describes a class of methods where the density is estimated for subsets of the given samples. The resulting distributions are compared to estimate which bandwidth should be most appropriate for the full set[92].

When using cross-validation, the allowable bounds for the bandwidth must be set. As bandwidth is comparable to variance (σ^2) for a Gaussian kernel, it is therefore a non-negative number. However, and especially for small sample sizes, allowing the bandwidth to approach zero may not be appropriate. If, as posited earlier, the unlikely circumstances unfolds where all of the scores are in perfect agreement, the resulting bandwidth would be nearly zero and the density estimate would be a single spike at the shared value. This goes against the aforementioned observation that there is some inherent uncertainty to the expert assessment and therefore should be some spread to the distribution. As such, a lower bound

of 0.05 is set for the bandwidth

Another feature of KDE is the application of relative sample weights. Each sample value is given a relative weighting towards how much it should influence the resulting distribution, as shown in Figure 5.2. This is important as it can be used to provide the capability described previously to apply a relative confidence to the opinions provided by experts. Additionally, a relative total weight for each aspect of fidelity can be applied. For a given method of approximating a system, the resolution, abstraction, and scope of the model each have some influence on predictive capability.

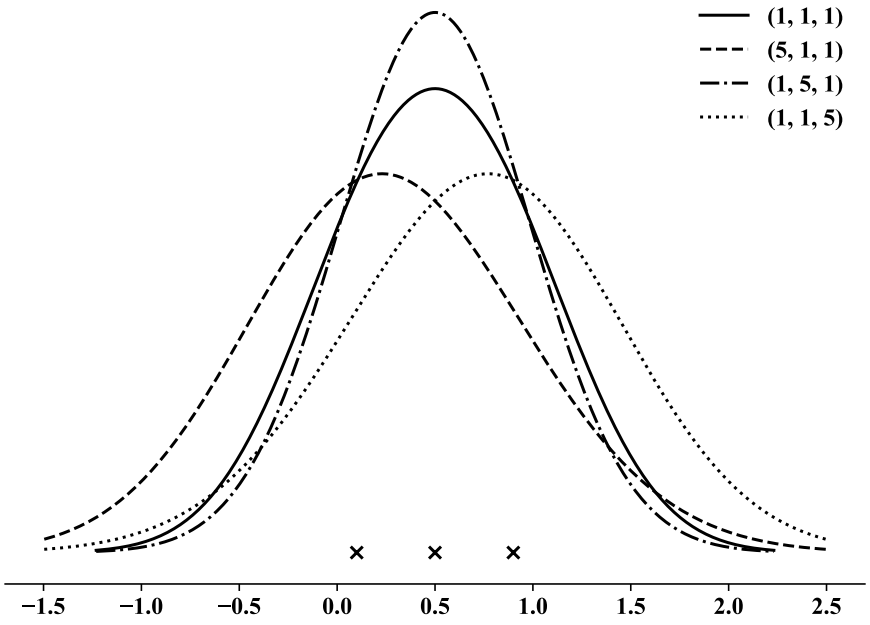


Figure 5.2: Weighted KDE: Different Weightings for Three Sample Values

The relationship between resolution, abstraction, and scope is complex, so additional research and the use validation data would be required to accurately predict their relative weightings for a given model or model set. In lieu of that level of information, the aspects of fidelity should be initially equally weighted. What follows is an example of how the relative sample weights for the three fidelity aspects is generated for the same and differing numbers of opinions for each. It should be noted that the full vector of sample weights is

normalized internally by the KDE code.

Number of opinions for each aspect	Relative sample weights
$[1], [1], [1]$	$[1], [1], [1]$
$[2], [1], [3]$	$[\frac{1}{2}, \frac{1}{2}], [1], [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$

Table 5.2: Sample Weights for Expert Fidelity Assessments

Kernel density estimation is a commonly used method in the machine learning community, and as such, there are a number of available implementations. As mentioned above, the implementation used in this work is the the one developed in the SciKit-Learn package[91]. It was selected since it is open-source, community-verified, and shown to be one of the more efficient implementations, at least within the Python language[92].

5.3 Assessment of Model Set 1: Notional Model Set

To further develop and evaluate a methodology, a set of four notional models is used. For each included dimension of fidelity, a model order is defined. Four different conditions are example to show the interplay between resolution, abstraction, and scope, and show the resulting KDEs:

1. All models increasing in resolution, abstraction, and scope
    - Resolution: [Model 1], [2], [3], [4]
    - Abstraction: [Model 1], [2], [3], [4]
    - Scope: [Model 1], [2], [3], [4]
2. All increasing, ignoring scope
    - Resolution: [Model 1], [2], [3], [4]
    - Abstraction: [Model 1], [2], [3], [4]
3. Increasing for resolution and abstraction, fixed scope

- Resolution: [Model 1], [2], [3], [4]
- Abstraction: [Model 1], [2], [3], [4]
- Scope: [Model 1, 2, 3, 4]

4. Increasing for resolution and abstraction, decreasing scope

    - Resolution: [Model 1], [2], [3], [4]
    - Abstraction: [Model 1], [2], [3], [4]
    - Scope: [Model 4], [3], [2], [1]

The resulting score values for the last case are shown in Table 5.3, and the density estimates are shown in Figure 5.3a, 5.3b, 5.3c, and 5.3d. The first case, shown in Figure 5.3a, describes the case that, moving from model 1 to 4, the models improve in resolution, abstraction, and scope. This is unrealistic, since, as described in Chapter 4, scope typically decreases as resolution increases and abstraction decreases.

Since scope is being introduced in the fidelity framework developed herein, the second case, shown in Figure 5.3b represents what would happen if it were ignored. The density estimates appaer identical to those of case 1, though, when scope is typically ignored, it is because all of the models have the same scope. Case 3 describes the case of including scope in the description and stating that all models have the same scope. Examining Figure 5.3c shows density estimates different from the previous two cases, since accounting for that fixed scope implies a similar for the models that was not accounted for in case 2. Even for this notional set of models, this shows the danger of leaving out an important characteristic of fidelity, as there is a significant shift in the estimated fidelity ranking.

Case 4 shows the more realistic alternative to case 1, as described above. The tanking with regards to scope follows the opposite pattern from the other two aspects, since scope is often decreased as resolution and abstraction improve. As a result, there is a great deal more overlap between the distributions, meaning that it is much less obvious that model 4 is the highest fidelity model since it has the worst scope of the group.

Table 5.3: Notional Model Fidelity Scores: Case 4

Model ID	Resolution Score	Abstraction Score	Scope Score
1	0.1	0.1	0.4
2	0.2	0.2	0.3
3	0.3	0.3	0.2
4	0.4	0.4	0.1

Visual assessment of the position and shape of these distributions can give some insight into the model set as it is currently understood. However, it can be difficult to glean much from this if the distributions are similar. This problem is exacerbated as the size of the model set rises. Additionally, one of the goals of this process is to generate an estimate of the probability that a model is the highest fidelity in the set. To generate model fidelity assessments as probability values, there must be a further understanding of what is actually being asked.

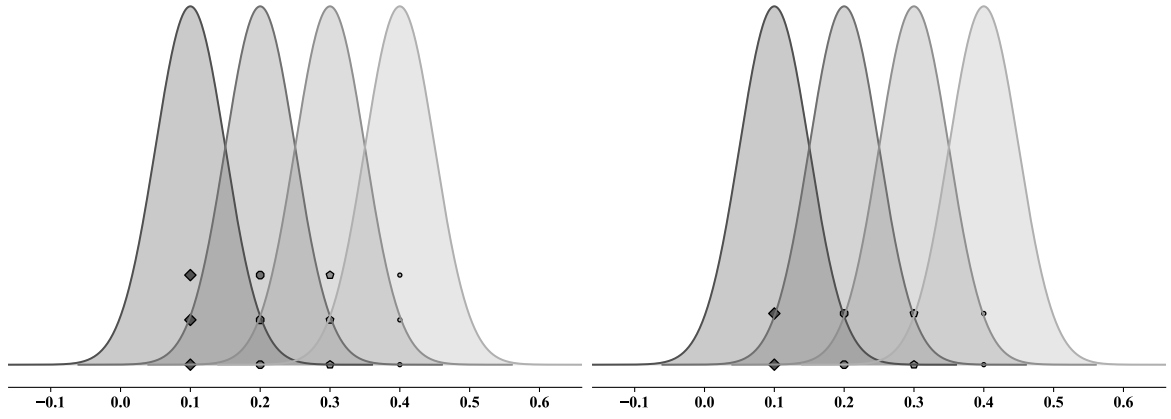
5.3.1 Calculating Probability of Highest Fidelity from Density Estimates

Pairwise Fidelity Comparison

In an ensemble learning method such as Bayesian model averaging, the model probability is technically the probability that a given model is the most likely to be able to generate the available relative truth data. As mentioned before, this is understood by the author to be an estimate of the probability that a given model is the highest fidelity in the current set. To estimate the probability that model X has the highest fidelity in the set is equivalent to saying that the fidelity score of X is greater than the fidelity score of model Y, Z, and all of the other models in the set. The probability that X has a higher fidelity than Y can be written as follows:

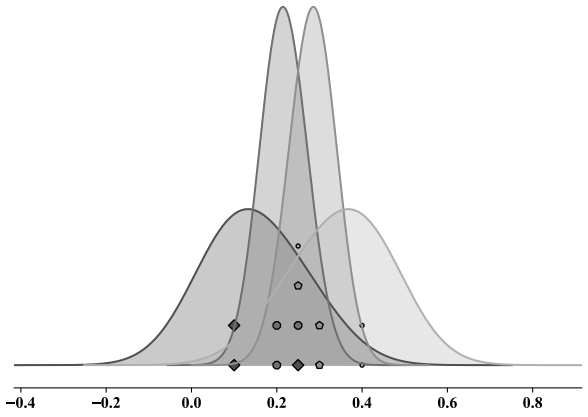
$$P(X > Y) = P(X - Y > 0)$$

Looking back at how kernel density estimation works, for corresponding samples, a distribution is placed with a mean of $X^{(1)}$ with bandwidth σ_X^2. Note that bandwidth is uni-

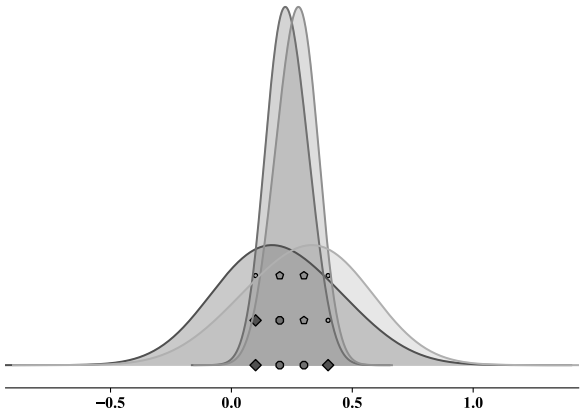


(a) All Increasing

(b) No Scope



(c) Fixed Scope



(d) Reversed Scope

Figure 5.3: Notional Kernel Density Estimates

formly defined for a given model. Since the kernel in used in this work is based on a normal distribution, this can also be written as $\mathcal{N}(X^{(1)}, \sigma_X^2)$. Similarly, the distribution generated for the corresponding value with respect to model Y can be written as $\mathcal{N}(Y^{(1)}, \sigma_Y^2)$. Since the response of interest is $P(X - Y > 0)$, the difference of X and Y is needed. It is known that the difference of two normally distributed random variables is a normally distributed random variable where the mean is the difference of the means and the variance is the sum

of the variances:

$$Z = X - Y$$

$$Z \sim \mathcal{N}(\mu_X - \mu_Y, \sigma_X^2 = \sigma_Y^2)$$

As such, the distribution of $X - Y$ can be found by evaluating a kernel density estimate using the difference of the sample values and using the sum of the bandwidths. Then, $P(X > Y) = P(X - Y > 0)$ then becomes the area under the positive section of the resulting density estimate. Correspondingly, $P(X < Y) = 1 - P(X > Y)$. The ***SciKit-Learn*** implementation returns the distribution curve as points on a line. The ***InterpolatedUnivariateSpline*** method in ***SciPy*** is used to spline the curve and allows for efficient and accurate integration of the area under the positive portion of the curve[93]. This is shown in Figure 5.4 and the difference in the score values betwen models 1 and 2 are as follows:

$$\overset{\text{Model 1}}{\begin{bmatrix} \text{Resolution}_1 \\ \text{Abstraction}_1 \\ \text{Scope}_1 \end{bmatrix}} - \overset{\text{Model 2}}{\begin{bmatrix} \text{R}_2 \\ \text{A}_2 \\ \text{S}_2 \end{bmatrix}} = \overset{\text{Difference}}{\begin{bmatrix} \Delta_R \\ \Delta_A \\ \Delta_S \end{bmatrix}}$$

Given this process, all of the pairwise probabilities in the set must be calculated, the number of which can be found as follows:

$$\binom{n}{2} = \frac{n!}{2(n-2)!}$$

For each pairwise combination of two models $P(i > j)$ and $P(i < j)$ can be found using the above-described process, re-written here:

1. Find difference in scores for models M_i and M_j
2. Define KDE with Δs and $bandwidth = M_i.bandwidth + M_j.bandwidth$

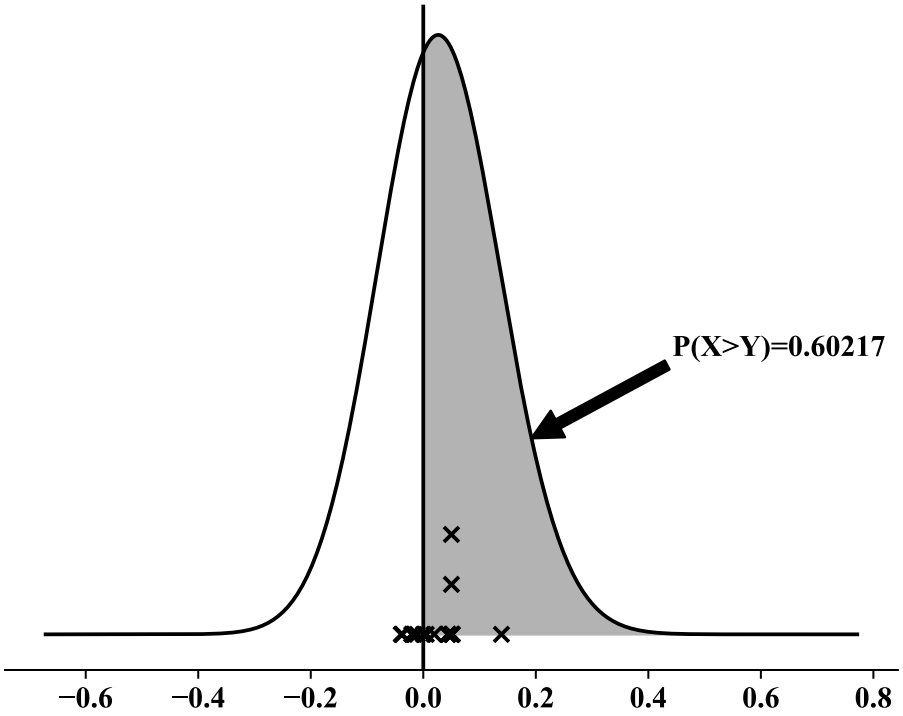


Figure 5.4: $P(X > Y)$ Given Difference in Sample Values

3. Spline KDE distribution using $SciPy.InterpolatedUnivariateSpline$

4. $P(M_i > M_j) = \int_0^\infty KDE_{i-j} df$

Model Probabilities in Set

Now that the pairwise probability between two models can be evaluated using the aforementioned method, the probability that a given model has the highest fidelity in its set can be calculated. For a model set of size m, described as $M_1, M_2, ..., M_m$, The probability that a given model's fidelity is the highest in the set is the proabaility that it is greater than each of the other models. This can be written as follows:

$$P(M_1 \text{ is best}) = P(M_1 > M_2 \text{ and } M_1 > M_3 \ldots M_1 > M_m) = P(M_1 > M_2 \cap M_3 \cap \ldots \cap M_m)$$

Since these probabilities are independent, the result is simply that product of all of the relevant pairwise probabilities.

$$P(M_1 \text{ is best}) = P(M_1 > M_2) * P(M_1 > M_3) \ldots * P(M_1 > M_m)$$

Writing this more generally leads to Equation 5.3.

$$P(M_i \text{ is best in set}) = \prod_{j \neq i}^{m} P(M_i > M_j) \tag{5.3}$$

All that is needed to perform this calculation is the pairwise probabilities between each model. For model 1, this results in the following:

$$\begin{aligned} P(M_1 \text{ is best in set}) &= P(M_1 > M_2 \cap M_3 \cap M_4) \\ &= P(M_1 > M_2) * P(M_1 > M_3) * P(M_1 > M_4) \\ &= 0.45443 \times 0.41406 \times 0.41390 \\ &= 0.077880 \end{aligned}$$

Once all of the probabilities are calculated, they are normalized so that the sum is one, as follows:

$$\begin{matrix} \\ P(M_1 \text{ is 1st}) \\ P(M_2 \text{ is 1st}) \\ P(M_3 \text{ is 1st}) \\ P(M_4 \text{ is 1st}) \end{matrix} \overset{\prod_{j \neq i}^{m} P(M_i > M_j)}{\begin{bmatrix} 0.07788013889103614 \\ 0.09363176097471354 \\ 0.15590443028968273 \\ 0.18735893370696982 \end{bmatrix}} / \sum \left[\prod_{j \neq i}^{m} P(M_i > M_j) \right] = \begin{bmatrix} 0.1512895905422784 \\ 0.18188861732047215 \\ 0.3028592110660459 \\ 0.3639625810712037 \end{bmatrix}$$

Given this method, other probabilities can also be found. If it is of interest, the probability that a model has the lowest fidelity in the group can be found fairly easily. It is simply the probability that a model is worse than all of the other models, and $P(X < Y) = 1 - P(X >$

Y). This is represented in Equation 5.4.

$$P(M_i \text{ is worst in set}) = \prod_{j \neq i}^{m} [1 - P(M_i > M_j)] \tag{5.4}$$

This could be extended again to find the probability that a particular model ranks second highest, or third, fourth, etc. This requires slightly more work, as

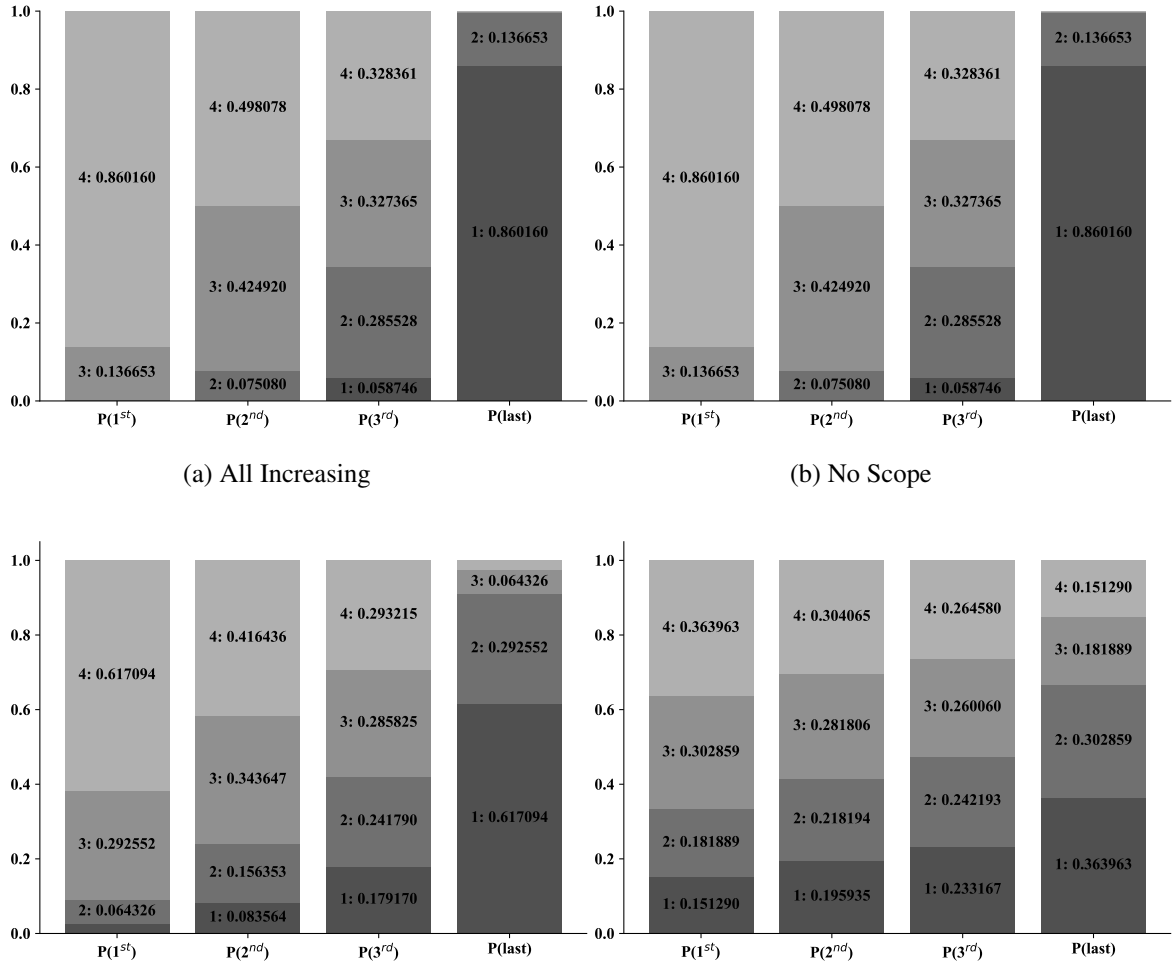
$$P(M_1 \text{ is } 2^{nd} \text{ in set of 3}) = P(M_1 > M_2 \text{ or } M_1 > M_3) = P(M_1 > M_2 \cup M_3)$$

and $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. As the number of models increases, the inclusion-exclusion principle must be used to expand the probabilities. The inclusion-exclusion in set theory describes the process for finding the union of sets:

1. Add the cardinalities of the sets: $P(M_1 > M_2) + P(M_1 > M_3) + P(M_1 > M_4)$
2. Subtract the pairwise combinations: $-[P(M_1 > M_2 \cap M_3)]$
3. Add the three-model combinations if appropriate
4. ... continue based on size of set

Including the probability sets up to the selected level, alternating between addition and subtraction, the probability of being at a certain level can be calculated. For the notional model set, the probability that each model's fidelity score ranks first, second, third, or fourth is shown in Figure 5.5a, 5.5b, 5.5c, and 5.5d.

Corresponding to the four cases described previously, if the scope is ignored because all of the models have the same scope, the probability model 4 is the highest fidelity is about 86%. If scope is considered and set equal, the highest fidelity probability for model 4 drops to 62%, a 28.3% error. This shows the significant difference that can occur if a fundamental characteristic of fidelity is left out. The probabilities are closer together when scope is included if the scope is fixed. If scope is ignored and the models follow the more



(a) All Increasing

(b) No Scope

(c) Fixed Scope

(d) Reversed Scope

Figure 5.5: Notional Probabilities of Highest Fidelity, 2nd Highest, etc. for Four Cases

realistic case, where scope is in the opposite order from resolution and abstraction, then the probability for model 4 is about 36%, a 57.7% error. Model 4 having the worst scope and model 1 having the best scope brings all of the density estimates closer, and, by extension, the probabilities are similarly valued. This leads to Observation 1.1.4 regarding Research Question 1.1.

Observation 1.1.4 *Leaving an important aspect of fidelity out of the framework can lead to over 50% error in the calculation of probability of highest fidelity in descriptive assessment of fidelity.*

Model probabilities can now be generated based on a number of appraisals based on the three aspects of fidelity. Additionally, the use of kernel density estimation to generate model probabilities leaves the door open to incorporate further assessments as other information becomes available. Importantly, this method of calculating the probability of highest through lowest fidelity only depends on the two-model comparisons in the model set, which, for four models, means twelve sets of calculations, and can be done a priori. Given this platform, generating fidelity scores based on available model data is the next problem to be addressed. However, before a method can be developed, actual model data is needed.

5.4 Model Set 2: I-Beam Modeling

5.4.1 Introduction

To further develop a methodology, data is needed, hence, a model set is needed. Drawing inspiration from a Federal Aviation Administration finite element modeling tutorial, there are four different ways that an simple I-beam can be represented in a finite element model[94]:

1. 1-D/Beam elements
2. 2-D/Shell elements for web, 1-D elements for caps
3. Shell elements for web and flanges
4. 3-D/Solid elements

The four options are also shown in Figure 5.6.

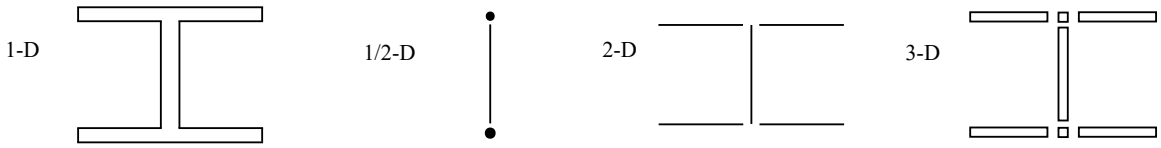


Figure 5.6: I-Beam Finite Element Representations

In Safarian's presentation, the solid elements were left out, but the included models were compared to a theoretical result in terms of deflection and stress, showing up to a 19% variation based on the element type used to represent the structure.

In addition to differences in shared responses such as displacement and stress, there are additional reasons why a different element type might be used. One-dimensional elements cannot represent anything about a specific point in the cross-section. Two-dimensional elements are not capable of showing how stress varies through the thickness of a flange or the web. Because of this, if the stress at a specific point on the surface of the beam is needed, a shell model can approximate it, but only a solid model can describe how the surface stress differs from the internal stress in the material. If the beam is to be affixed to another structure using bolts through particular locations on the flanges, one of these higher resolution models (shell or solid) would be required to represent how the beam structure would be affected.

To represent a structure using finite elements, a mesh must be overlaid on the geometry to define zero-dimensional node locations and the elements that connect them. For complex geometries, it is sometimes easier to define a mesh using a graphical finite element pre-processor, so the mesh can be visually inspected. Automation becomes tricky, as a badly formed mesh can skew the results. This occurs when three-sided elements are not equilateral and four-side elements are not square. However, an I-beam represents a simple geometry, which makes the mesh generation process easier. Additionally, even though solid elements require more computational effort than beam elements, since there are additional degrees of freedom in the element that can deform, none of these models are particularly computationally intensive due to the geometric simplicity. Despite the ease of generation and execution, since they can be represented a number of different ways, and, as the cited work shows, can show some variation due to the method of representation, they present a valid set to be used as a canonical example to further develop methods for understanding multifidelity model sets.

5.4.2 Description of Structure

For a fixed cross-section, as the length is varied, the accuracy of different mathematical structural representations is affected based on its inherent assumptions. Specifically, as the order of magnitude of length approaches that of the cross-sectional dimensions, e.g. height and width, the structure can no longer be described as "beam-like." Stating that a structure is a beam implies that it is much larger in one dimension than in the other two. Certain mathematical representations of the behavior of beam-like structures rely on this to remain valid. Because of those cases, it is known that a two-dimensional, or shell, representation should be more generally applicable since it allows for cross-sectional effects. In the case where the behavior is dependent on the behavior through the thickness of the material, the three-dimensional, or solid, elements are, in general, even more accurate.

For this model set, a right-angled I-beam cross-section similar to an AISC W5x16 section is used, as shown in Figure 5.7. The dimensions are shown in Table 5.4. The material properties are that of a generic steel, defined in Table 5.5. The attribute to be varied, as mentioned above, is the length of the beam. The length is varied such that the ratio of length to cross-section height goes between 2 and 40. This results in lengths between $\approx$10 and $\approx$200 inches. This is selected as mentioned above to push the limits of the modeling assumptions for the selected element types. From Roark's formulas for stress and strain, beam assumptions are valid under the following conditions of span over depth[95]:

1. span/depth $>= 8$ for metal beams of compact section
2. $>= 15$ for beams with relatively thin webs
3. $>= 24$ for rectangular timber beams

Considering an I-beam as a relatively thin-walled cross-section, the second condition is the best fit. Because of this, the accuracy of the one-dimensional numerical approximations is exptected to take a hit below $15*$length / height≈ 75 inches.

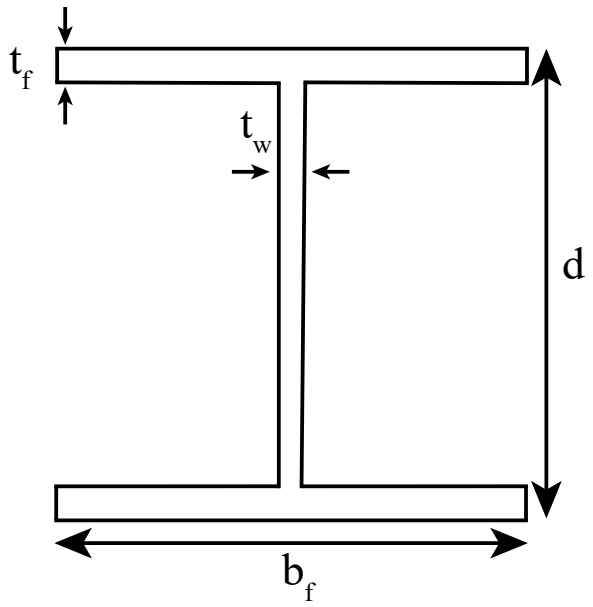


Figure 5.7: I-Beam Section

Table 5.4: AISC W5x16 Dimensions[96]

Dimension	Value (Imperial)	Metric
d	5.01 in.	12.7254 cm
b_f	5 in.	12.7 cm
t_f	0.36 in.	0.9144 cm
t_w	0.24 in.	0.6096 cm

As stated previously, this model was specifically selected such that the geometry is relatively simple: it is simply a projected cross-section. Python code was developed to define the grid points of a finite element mesh based on the selected dimensionality: 1-D, 1/2-D, 2-D, or 3-D. Upon initial generation, there is only assumed to be a single finite element in any given direction. A 1-D model assumed to be a single beam from end to end. A shell model has a single element for each of the four flanges, and one for the web. A solid model has one element for each of the flanges and one for the web, similarly to the shell models, but since thickness is included, additional elements are placed at the upper and lower intersections of flange and web. The node locations and mesh connectivity is automatically calculated using simple geometry, and must be used to create a finite element input file.

Table 5.5: Steel Properties[95]

Property	Value	Unit
Density (ρ)	7.85	g/cm^3
Modulus (E)	210	GPa
Poisson ratio	0.3	-

5.4.3 Finite Element Solver: MSC Nastran

The finite element software used in this work is MSC Nastran, specifically version 2017.1[97]. It is an industry standard finite element software package that provides a great deal of capability that has been developed and maintained since its initial inception as NASA software in the 1960s.

Nastran input files are text-based and formatted similar to legacy Fortran punch-cards. Each line is 80-characters, and those characters are broken up into eight or sixteen-character fields for input data. Typically, a mesh is generated using a graphical pre-processor that writes the appropriate input file. However, and especially in the case of such a simple geometry, it is often easier to define the attributes of the input file directly. Despite this, the specific formatting required for a Nastran input file can often be an impediment. For this reason, the author has developed a code over the past few years referred to simply as "Nastran utils." More information regarding the format of Nastran cards can be found in the Quick Reference Guide[98].

5.4.4 Python Package: Nastran Utilities

Nastran utils, or *nastran_utils*, is a Python package developed by the author, for the purpose of reading, writing, and modifying Nastran entries, called cards after the original Fortran implementation, as Python objects. Cards can be defined by reading an existing input file or directly in Python, and the code handles conversion to a formatted string that Nastran will accept.

The Nastran utility package is currently in version 0.6, and has been developed for

Python 2.7+ and 3.5+ compatibility. It can parse Nastran short, long, and comma-delimited format, and by default will write out cards in the format that retains the most precision. However, it can be set to write out in a certain format, since certain add-ons can only parse the short, 8-character, format. Due to the character limits for each field, certain tricks must be utilized to retain precision, including how Nastran scientific notation only requires a sign with no "e." Additionally, trailing zeros will be removed if they fall after a decimal.

Over 150 bulk data entry classes are included, as well as over 30 case control command classes. When reading a file, cards that are not yet implemented will be written to a log so that the user can determine what to do with those entries. Certain entries, such as materials and properties, have formatted comments that include meta-data such as the names of the entries. This information can also be parsed and written back out.

Nastran models are defined appropriately using the *NastranModel* class, which includes all of the entries, solution information, etc. When writing an input file, all of the entries are put in a specific order and keywords are added so that Nastran can use the file and the file is as readable as possible. There is an *analyze* method which can call Nastran with a given file. This tracks the solver process on the machine, can kill it if it seems to be taking too long, and will check the *f06* output file to see that the model ran successfully, or raise the appropriate Python exception based on a Nastran model failure.

The mesh entries, loads, boundary conditions, material properties and supplemental information such as solution parameters, are defined via Nastran utils, and then written to a Nastran input file, typically called a bulk data file or bdf. The ".bdf" file extension (".dat" is also accepted) can then be processed and analyzed by the MSC Nastran solver.

5.4.5 Model Types

The primary difference in model types for this model set is the type of finite element. The types fall into the four previously defined categories shown in Figure 5.6, based on element dimensionality:

- 1-D: The cross-section is represented as an element property and only the length is visible
- Hybrid 1/2-D: Tranverse web represented with shell elements, but flanges represented with attached 1-D elements
- 2-D: Also referred to as shell elements, where the cross-section is explicitly defined but the thickness of each section is simplified
- 3-D: Length, cross-section shape, and thickness are all explicitly represented

Due to the simplicity of the geometry, the scope is fixed for all models, but the resolution and abstraction changes with the selected element type. There are multiple types of elements for each of the above categories, as listed in Table 5.6. Much of the following information is from the MSC Nastran Linear Static User's Guide[99], Quick Reference Guide[98], some of which can also be found on the MSC Software blog[100].

Rods are the easiest elements to use, as only an area needs to be provided to define the cross-section. However, this means that the shape of the cross-section is not represented, and Rod elements are only for carrying axial and/or torsion loads. Because of this, Rod elements are not used on their own in this model set, though they are included in the 1/2-D hybrid models. Bars and Beams do require some understanding of the cross-section. Beam elements are able to represent taper and warping effects that a Bar element cannot. Even though Beam elements are more capable than bar elements, in a simple case such as this, where the cross-section is symmetric and the beam isn't tapered, they should behave similarly.

The compatible combinations of elements is listed in Table 5.7. As stated above, Rod elements are not included on their own, while Bar and Beam elements are. Additionally, it should be noted that there are other types of shell and solid elements to account for triangular or oddly shaped sections, but they are excluded here since the geometry has only right angles.

Table 5.6: Nastran Element Types

Dimensionality	Element
1-D	Rod
-	Bar
-	Beam
2-D	Quad4
-	Quad8
-	QuadR
3-D	Hex

Table 5.7: Model Set 2

ID	Element Selection
1	Bar
2	Beam
3	Quad4/Rod
4	Quad4/Bar
5	Quad4/Beam
6	Quad8/Rod
7	Quad8/Bar
8	Quad8/Beam
9	QuadR/Rod
10	QuadR/Bar
11	QuadR/Beam
12	Quad4
13	Quad8
14	QuadR
15	Hex

5.4.6 Problem Definitions

Definition of a set of models is dependent on a defined problem. In this case, three problems are defined to test the capabilities and limitations of the canonical model set. Comparing different model responses also allows for verification of the model generation processes. The responses for each of the three problems are written out to a "hierarchical data format," specifically an *HDF5* binary file[101]. This filetype was added as an output option for MSC Nastran as of version 2016 as it is a standardized binary data format that is efficient and retains a high numerical precision[102]. For reference, most of the other Nastran binary

output file formats are specific to Nastran. By using a standard data format, the outputs can be parsed using an open-source Python package *h5py*[103]. Other languages also have packages to work with this format, and there are graphical interface packages for examining the data. The *h5py* package was used to automate extraction and processing of the relevant outputs of each response scenario. Also, for more information regarding the following solutions in Nastran, see the MSC Nastran linear analysis user manual[99].

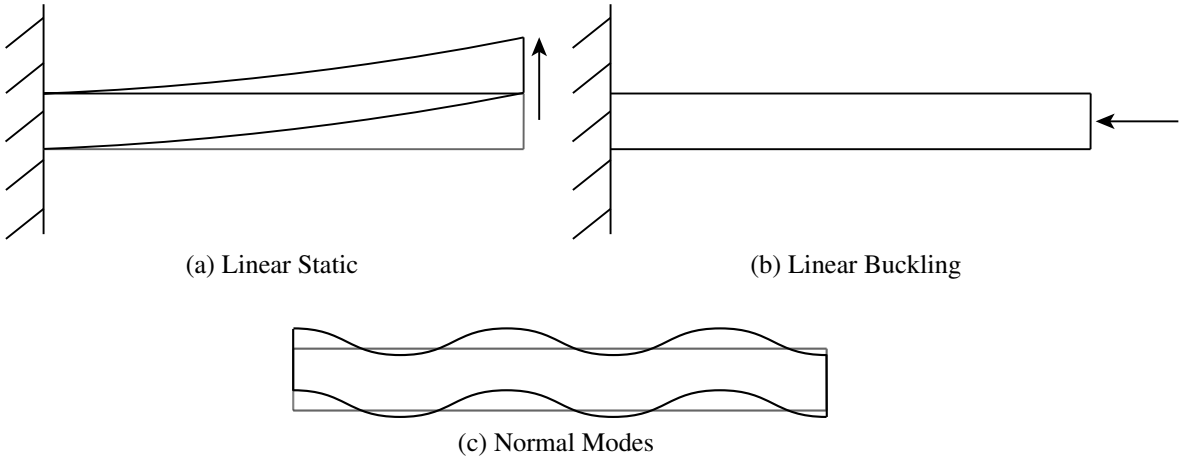


Figure 5.8: Beam Problem Definitions

Linear Static

The first defined problem is the linear static deflection of a cantilever beam under a transverse tip load, as shown in Figure 5.8a. In Nastran, linear static analysis refers to solution 101[98]. This is a very common and simple problem scenario for a beam and can therefore be used for verification of the models. One of the model is fixed and a 10,000 newton or 10 kilonewton transverse load is applied to the other end. The load is oriented to be applied along the typical vertical load-bearing direction of an I-beam.

For shell and solid elements the flanges can displace separately from the web, so care must be taken when applying a load to the free face. The load is applied at a single grid point offset slightly from the face, and the load is transferred only to the grid points of the transverse web, as they are aligned with the direction of the force. This load transfer is done

internally to Nastran through the use of a Nastran *RBE3* element. While the *RBE* in *RBE3* stands for rigid-body element, an *RBE3* is actually an interpolation constraint element that "defines the motion at a reference grid point as the weighted average of the motions of a set of other grid points[98]."

The response that is recorded is the deflection of the beam tip under load. Displacement is reported by node or grid point, so shell and solid elements have multiple grid points at the loaded end. Because of this, the displacement vectors are gathered for all of the nodes on the tip face, and the average vector is found. Whether developed from a single grid point or many, the response that is saved is the magnitude of the resulting vector.

Linear Buckling

The second case is the linear buckling critical eigenvalue for a cantilever beam under a compressive tip load using Nastran solution 105, as shown in Figure 5.8b. Similarly to the linear static case, one end of the beam is fixed, and a ten Kilonewton load is applied to the opposite end. Unlike in the transverse deflection case, the load is being applied along the axial direction of the beam, so it can simply be transferred to the entire face.

Once the linear buckling analysis is performed, one of the responses is a set of eigenvalues representing the linear buckling modes of the structure. The critical eigenvalue is the first value in this series, and is a scaling factor on the load to represent where the structure is predicted to buckle. This response is recorded for each linear buckling model.

Normal Modes

The third modeling problem definition is the first non-zero eigen-frequency for a beam in free vibration using the normal modes analysis of Natran solution 103, as shown in Figure 5.8c. No loads or boundary conditions are applied to the structure, and the resonant frequencies and shapes of the structure. Since the structure is allowed to vibrate freely, the first few modes that are returned are rigid-body rotations, so there is no relative displace-

ment within the structure itself. The recorded response is the frequency, in Hertz, of the first non-trivial mode, when the structure vibrates with some deflected shape.

5.4.7 Design of experiments

Given the variables and ranges of interest for a particular problem, a design of experiments must be generated to efficiently explore the design space. Given that the length of the beam is the only variable, a simple method could be used to find a number of points to evaluate over the range of $\approx$[10, 200] inches. However, considering that most problems will have more than one variable, a more general process is put into place in the code to find a design of experiments. First, points at the lower bound, upper bound, and center are selected via a three-level full-factorial design. Second, a maximin Latin Hypercube design with 50 points is generated to fill the rest of the space. This 53-point design is applied to all of the model types and problem definitions. These designs are generated using the *pyDOE2* Python package[104].

5.4.8 Mesh Convergence

As mentioned above, the mesh for a given model instance is generated initially with one element in each available direction. However, under most circumstances, the density of the mesh must be increased for the model to provide its best prediction of the system response. As such, the process of mesh convergence must be undertaken to some extent to provide a reliable predictive platform. Mesh convergence can often be a very difficult and laborious process for models with complex geometry, but the relative simplicity of this model set allows for a fairly thorough exploration. Converging on a mesh requires multiple calls of the finite element solver, tracking the mesh density in each direction and the history of the predicted response.

Ths basic convergence criterion examines the last three mesh instances and compares their responses. If they are within a certain tolerance of each other, for this case study

0.5% was used, then the first of those three densities can be called sufficient. For a one-dimensional element type, the process is comparatively simple, as there is only one way to increase the density of the mesh. The density is simply increased by adding a single element axially until convergence is achieved.

For shell and solid elements, the density can also be increased across the cross-section. Specifically, the number of elements along each flange or web can be incremented. For these types of models, the convergence process is therefore iterative. Additional criteria are added regarding each sub-convergence to avoid divergent behavior. Specificaly, if the response has changed a significant amount or the density in that direction has been increased more than a certain amount, the process should stop and switch to the other direction, whether axial or through the cross-section. If the response has changed more than 25%, or the mesh density increased 20 times axially, or 5 times laterally, without convergence, then the process is interrupted and the direction changed.

Given the fifteen different element options, three different response scenarios, and 53-point design of experiments, mesh convergence was performed for each of the $2,385$ models. As mentioned previously, the convergence process requires multiple analysis calls. Specifically, in the case that the initial mesh density is sufficient, three calls are still needed to verify this sufficiency. This amount of analysis calls would be prohibitive for a complex model, but this model set was specifically selected due to the relatively small number of degrees of freedom. Additionally, due to the automation of the process, most of the models can be generated, analyzed, and post-processed in a second or less. As such, all of the models were generated and the responses saved on a single standard desktop over the course of about a week. The process could be distributed to multiple computers or even a remote distributed cluster if this time needed to be further reduced. Given a set of models and the corresponding responses, the data can be used to further develop the fidelity assessment methodology through comparative data analysis.

5.4.9 Descriptive Fidelity Assessment

Descriptive Orders

Prior to examination of any results, the models can be assessed using the descriptive approach derived in Section 5.2. The orders used to describe the relative resolution, abstraction, and scope of the fifteen models are shown in Table 5.8. The models are ordered for resolution, abstraction, and scope. Models that are similar with respect to that fidelity aspect are grouped together, such as [Bar, Beam] in Resolution (1). Since resolution and scope improve as they increase and abstraction improves as it decreases, the order could become confusing if it must be reversed based on the direction of improvement. Because of this, the orders should simply be provided from worst to best.

The orderings in Table 5.8 were generated by the author with the knowledge of the models in question to represent a sample descriptive assessment. Note that two different opinions were given for both resolution and abstraction. This represents the case that experts cannot arrive at a single agreed-upon representation of the model ordering in terms of those characteristics. Specifically for abstraction in this problem, a Quad4/Rod and a Quad8/Rod model should be very similar. However, since a Quad8 element is known to be more accurate than a Quad4[99], Abstraction (2) puts the Quad8/Rod model in a separate group, higher the Quad4/Rod model. This shows the flexibility of this method, using KDE to combine not only the fidelity aspects, but the opinions of multiple experts.

When there is disagreement, not only can all of the expert-provided orderings be used, but it is recommended. Including a case where Quad4/Rod and Quad8/Rod are in the same group, as in Abstraction (1), as well as one where they are in a separate group, as in Abstraction (2), will lead to a different result than if only one was included. If only Abstraction (1) was included, the two models would be given the same abstraction score. If only the Abstraction (2) were provided, it would not represent the similarity between the two models. Including both orderings results in abstraction scores for these two models that

Table 5.8: Initial Model Set 2 Assessment by Ordering

Fidelity Types	Model Ordered Groups
Resolution (1)	[Bar, Beam] [Quad4/Rod, Quad8/Rod, QuadR/Rod, Quad4/Bar, Quad8/Bar, QuadR/Bar, Quad4/Beam, Quad8/Beam, QuadR/Beam], [Quad4, Quad8, QuadR], [Hex]
Resolution (2)	[Bar, Beam], [Quad4/Rod, Quad8/Rod, QuadR/Rod], [Quad4/Bar, Quad8/Bar, QuadR/Bar], [Quad4/Beam, Quad8/Beam, QuadR/Beam], [Quad4, Quad8, QuadR], [Hex]
Abstraction (1)	[Bar], [Beam], [Quad4/Rod, Quad8/Rod, QuadR/Rod], [Quad4/Bar, Quad8/Bar, QuadR/Bar], [Quad4/Beam, Quad8/Beam, QuadR/Beam], [Quad4, Quad8, QuadR], [Hex]
Abstraction (2)	[Bar], [Beam], [Quad4/Rod], [Quad8/Rod], [QuadR/Rod], [Quad4/Bar], [Quad8/Bar], [QuadR/Bar], [Quad4/Beam], [Quad8/Beam], [QuadR/Beam], [Quad4], [Quad8], [QuadR], [Hex]
Scope	[Bar, Beam, Quad4/Rod, Quad8/Rod, QuadR/Rod, Quad4/Bar, Quad8/Bar, QuadR/Bar, Quad4/Beam, Quad8/Beam, QuadR/Beam, Quad4, Quad8, QuadR, Hex]

shows that the Quad8 elements in the Quad8/Rod should be more accurate than Quad4's in Quad4/Rod, but that Quad4/Rod and Quad8/Rod should score more similarly than, say, Quad4/Rod and Quad4.

As mentioned previously, the scopes of all of the models are the same by definition, so there is one definite ordering for scope. This helps to point out how the method developed here can handle a different number of opinions for each aspect of fidelity. If the total score for each aspect is represented as 1, the single ordering for scope would be given a weighting

of 1 by defaut, while the two orders for resolution would be given weightings of 0.5 and 0.5. The same can be said for the two abstraction orders.

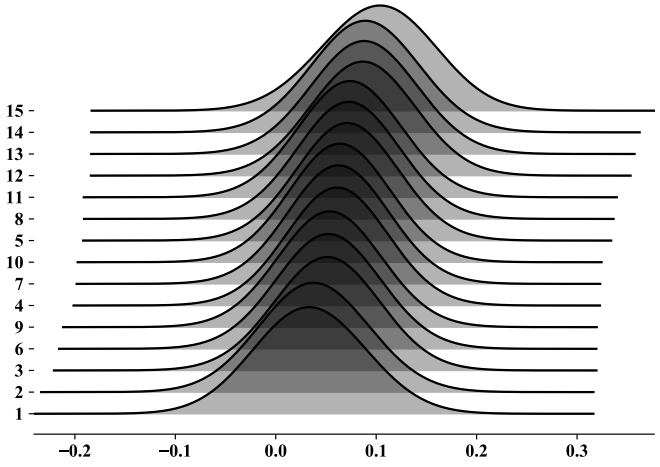
It is mentioned that this is the default behavior since if multiple opinions are given, and there is more confidence towards one over another, it could be given a higher relative weighting. For example, Resolution (1) could be given a weighting of 0.8, and (2) a weight of 0.2, instead of an even weighting, if it is believed Resolution (1) is more likely.

Additionally, if there was some justification based on experience that for the particular problem at hand that, say, resolution is more important than abstraction or scope to the fidelity of the models, then the total weighting of resolution could be increased with respect to the other two aspects. For example, resolution could be given a total score of 2, abstraction, 1, and scope, 1. By default, this means Resolution (1) would be given a weighting of 1, as would Resolution (2), meaning that these orders are each as important to the resulting KDE as the total weight for scope. Looking back at Figure 5.2 for how weightings can change the resulting distribution, the KDE would be shifted toward the score of the more highly weighted attribute.

Weighted KDE is a relatively new feature to the SciKit-Learn implementation as of the writing of this document. One of the benefits of the open source nature of the SciKit-Learn package, as with most Python projects, is that features can be tracked and more information can be found if desired. Specifically, weighted KDE can be traced to issue #4394 on the SciKit-Learn GitHub site.

Descriptive Fidelity Results

The initial density estimates and model probabilities based on the orders in Table 5.8 are in Figure 5.9 and Figure 5.10 respectively. The distributions are sorted by median instead of ID to show where each of the models fall in the set. When sorted by median, the models are in the same order as "Abstraction (2)" in Table 5.8: Bar, Beam, Quad4/Rod, Quad8/Rod, QuadR/Rod, Quad4/Bar, Quad8/Bar, QuadR/Bar, Quad4/Beam, Quad8/Beam,



(a) KDE

Median	Model	ID
0.03416	Bar	1
0.03829	Beam	2
0.05108	Quad4/Rod	3
0.05238	Quad8/Rod	6
0.05374	QuadR/Rod	9
0.06108	Quad4/Bar	4
0.06246	Quad8/Bar	7
0.06385	QuadR/Bar	10
0.07118	Quad4/Beam	5
0.07256	Quad8/Beam	8
0.07392	QuadR/Beam	11
0.08643	Quad4	12
0.08780	Quad8	13
0.08911	QuadR	14
0.1021	Hex Solid	15

(b) Median Values

Figure 5.9: Model Set 2 Descriptive KDE

QuadR/Beam, Quad4, Quad8, QuadR, and Hex.

Looking at the initial estimated model probabilities shows that, as expected, the solid model (15) scores highest. This is followed by the purely shell models (12, 13, and 14), followed by the QuadR/Beam model. Conversely, the one-dimensional element models rank the lowest (1 and 2), followed by the hybrid models with rod elements (3, 6, and 9).

Even if no model data was available, this still provides an understanding of the model set based on an intuitive and traceable process, which leads to Conclusion 1.2. What must be kept in mind, however, is that the relative magnitudes between the models is still comparatively notional until model data can be used to scale the relative magnitudes.

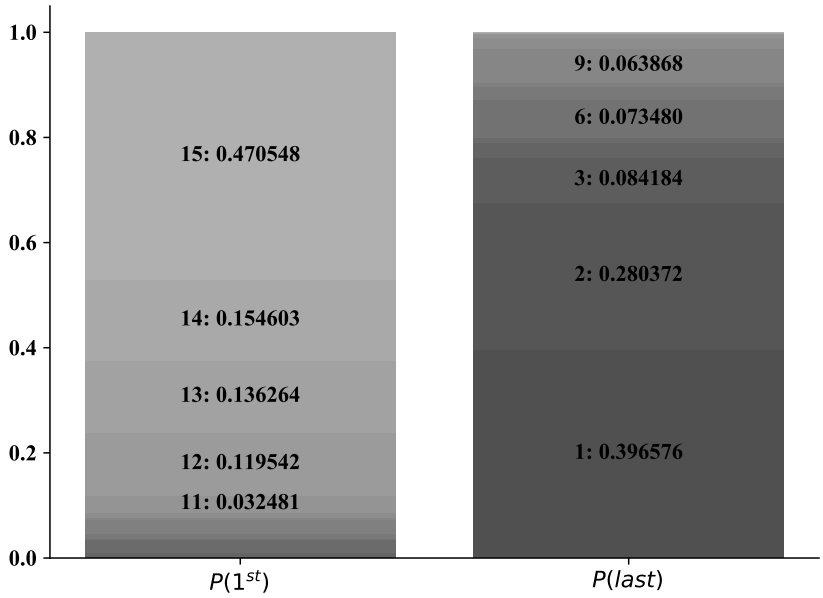


Figure 5.10: Probabilities of Highest and Lowest Fidelity for Full Model Set 2 From Descriptive Assessment

Conclusion 1.2 *Using resolution, abstraction, and scope to decsribe and compare the relative fidelity characteristics of models, and utilizing weighted KDE to combine opinions and represent the uncertainty related to that assessment, distributions can be generated to define a relative understanding of fidelity with respect to models in a multifidelity set. These distributions can then be used to calculate the probability of highest fidelity for each model, which both captures expert opinions and increasing model understanding, even in the absence of model predictions.*

5.4.10 Initial Data Examination

Now that an initial fidelity assessment has been performed, the available data may be taken into account. The results of the linear static, linear buckling, and normal modes responses are shown in Figures 5.11, 5.12, and 5.13. A number of initial observations can be made based on this data.

As expected, in Figure 5.11, the linear static tip deflection increases as the beam length

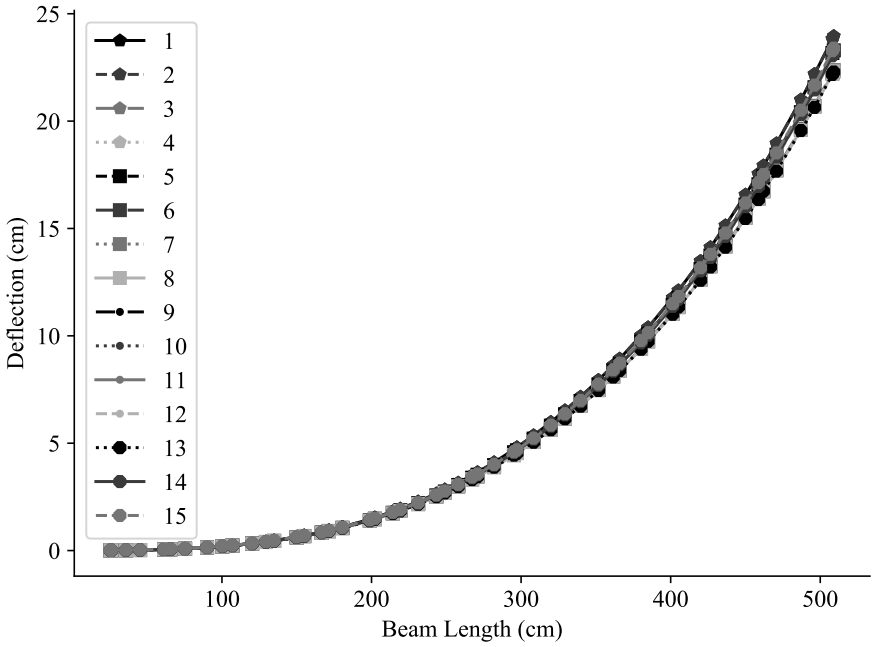


Figure 5.11: Model Set 2 Linear Static Results For All 15 Models

increases. Additionally, all of the models agree with each other, which helps to verify that the process of model generation, mesh convergence, analysis, and post-processing is working as expected. However, all of the models agree with each other very closely, which does not agree with the expected characteristics of divergence as the ratio of length to depth falls below a certain value.

Divergence for low length values does, however, show up in Figures 5.12 and 5.13. For the linear buckling results in Figure 5.12, there are four different groupings of results as the beam length decreases. This means that there is some agreement between models, but a clear difference between the way the beam is represented for low lengths.

The solid and three pure shell models all follow a trend towards a much higher critical load than the other models. The 1-D element type models flatten out around 100 cm, most of the 1/2-D hybrid topology models predict lower values, even though there is still a changing derivative. The Quad/Rod models in their current state, however, predict a nearly zero buckling load across the design space, which will be discussed further in a following

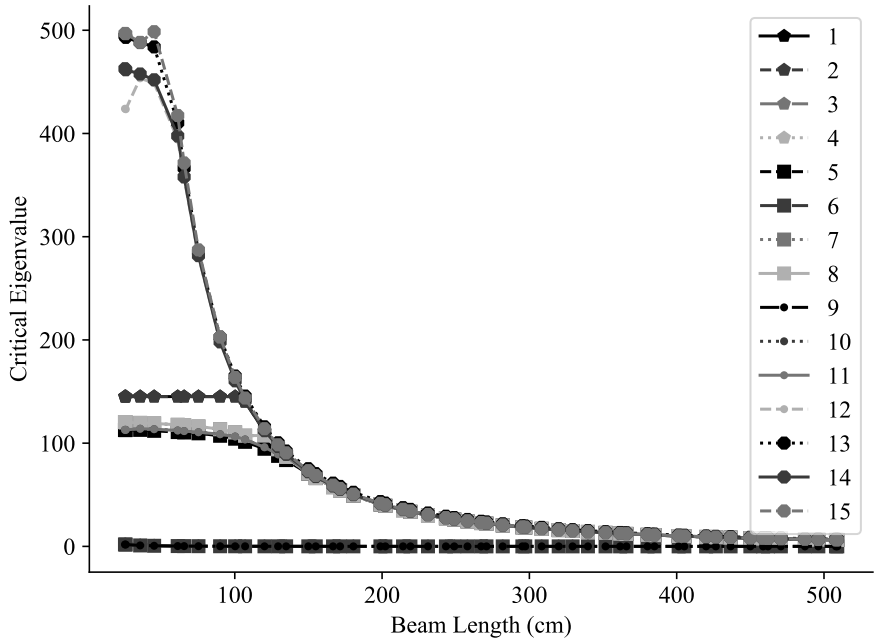


Figure 5.12: Model Set 2 Linear Buckling Results For All 15 Models

section. The normal modes models represent a third scenario. The Bar model diverges for low lengths, but the other models predominantly agree with one another, similarly to the linear static case. Due to the multiple distinct trends, including one at nearly zero across the design space, the linear buckling results are used for the initial development of the comparative data analysis-based fidelity scoring methods.

5.5 Fidelity Assessment Through Comparative Data Analysis

5.5.1 Research Question 1.3

Now that fidelity can be assessed based on expert opinion and a multifidelity data set has been developed, as described in the previous section, the use of available data can be addressed. The density estimation process is flexible enough to easily incorporate additional fidelity scores, so the question becomes:

Research Question 1.3 *How can the probability of a model being the highest fidelity be updated and corrected based on available model data?*

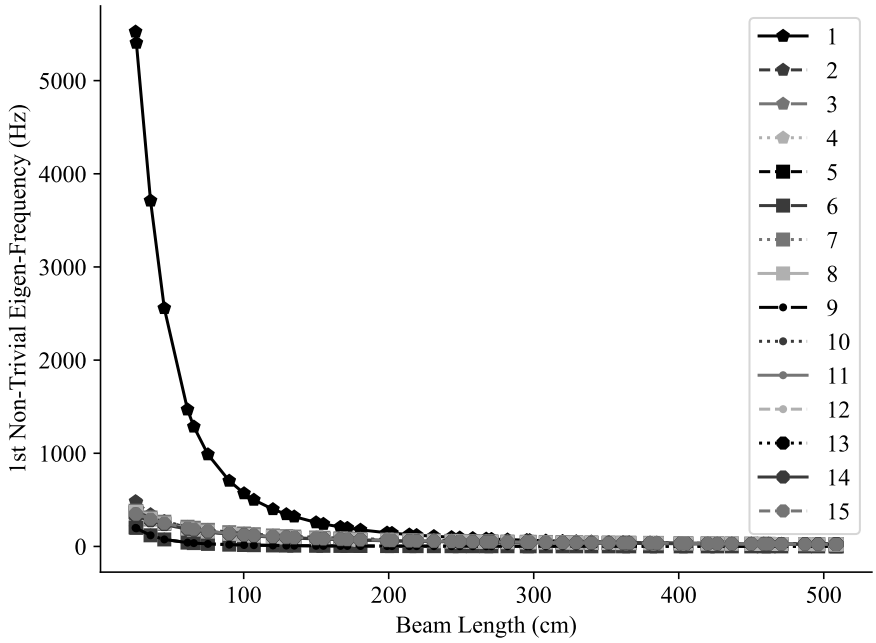


Figure 5.13: Model Set 2 Normal Modes Results For All 15 Models

5.5.2 Multi-Model Data Comparison

A multifidelity model set is essentially a number of different paths towards the same end. The main difference is that multifidelity models don't typically arrive at exactly the same value, which is why multiple options must be considered. As discussed in chapter 3, it is very difficult to find out which of the possible models generates the most objectively true result. Preferably, this would be done by comparison to real-world historical or experimental validation data, but, as discussed previously, this is often not generally available. However, even validation data is subject to the problems inherent in benchmarking.

Benchmarking, as discussed in Section 3.1, is the process of stating that one set of data is to be treated as the most accepted set, and variation from those values in a reference model should be treated as error. This process, is, however, always going to involve some level of subjectivity, as, by definition, there is no higher level of data to use for comparison.

The user in this case is presented with a multifidelity model set in the absence of validation data. It is presumed that the models in the set are of varying fidelity, meaning some

should be trusted more than others. That being the case, as model data becomes available, some information can be found by hypothesizing that each model is the most trusted, and using that data set as a benchmark. If this process is repeated for each model in the set, an approximation of a validation process can be performed by comparing the spread of results.

As a thought experiment, if a set of 100 models is being examined, and the responses for 99 of those models show some quantitative level of agreement, it is much more likely that the 100th model is a poor choice than that the other 99 are wrong. If the other 99 are wrong, then the experts developing the model set have put forth 99 inferior and 1 that is worth considering. Note that this example is for a very large model set. If there are only two models in the multifidelity set determining which one has a higher fidelity based only on the two data sets is essentially equivalent to flipping a coin. However, this is why the descriptive assessment of fidelity should be performed first, as it allows a chance for experts to record which models they believe to be valid based on past experience and an understand of the problem and model characteristics.

For this, it must be kept in mind that a likely follow-up to multifidelity model selection is the use of a multifidelity surrogate-based optimization technique. Multifidelity Gaussian process regression, also known as Co-Kriging, is commonly used in this process, as mentioned in Section 3.5. This is important, as it provides insight into how the quality of a multifidelity down-selection can be judged. The work of Toal discusses the development of a set of best practices for model selection in multifidelity regression[105]. That work is based on observations that while Guassian process regrssion is a powerful method for fitting a variety of data sets, they are subject to a type of overfitting. Depending on the selection of models, the response of the regression can be overconfident about the behavior of the system. Three primary rules were generated through examination of the literature, and statistical analysis of actual data sets. Experimental data was used to prove out the validity of these rules.

One of the rules of thumb states that the ratio of training points between low and high

fidelity should favor the model deemed as low fidelity. The main provision of this rule relates to the assumption that a higher fidelity model is more expensive than a lower fidelity model. A higher fidelity model should provide more intuition regarding the magnitude and trend of the data as opposed to covering the design space. This is an important consideration, but as it pertains primarily to required effort, is not as directly applicable to the fidelity assessment process developed in this section.

The other two rules speak to how the trends of the data sets compare to one another. If models generate data that are in agreement, or well correlated, they were able to prove that the resulting multifidelity interpolator will be more reliable. The specific metrics seleted to prove this fall into the category of regression goodness-of-fit metrics, specifically coefficient of determination, R^2, and the root mean square error (RMSE).

These goodness-of-fit metrics are assessed on pairwise combinations of data, typically used to verify quality of a regression with respect to the traininig or full data set. The coefficient of determination of determination is a measure of correlation and describes the amount of the variance of the truth data that the regression can predict, and is a value less than or equal to one, where a one represents a perfect fit. The root mean square deviation or error examines the magnitude of the error in the residusla between the data sets and provides a non-negative score. A value of 0 would reflect a perfect fit, but it nearly impossible in a practica sense. The R^2 and RMSE are defined as in Equations 5.5 and 5.6. The calculation of these metrics is performed using the metric scoring methods *r2_score* and the square root of *mean_squared_error*[91].

$$R^2 = \left(\frac{\sum_{i=1}^{n}(y_{e_i} - \bar{y}_e)(y_{c_i} - \bar{y}_c)}{\sqrt{\sum_{i=1}^{n}(y_{e_i} - \bar{y}_e)^2}\sqrt{\sum_{i=1}^{n}(y_{c_i} - \bar{y}_c)^2}} \right)^2 \tag{5.5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{e_i} - y_{c_i})^2} \tag{5.6}$$

Essentially what this is saying is that if nothing else a combination of models with a

high correlation score would make for a good selection in multifidelity refression, which should not be overlooked. For this work, this is logically extended to the above comments regarding benchmarking and model agreement. If approximations based on different sets of logical steps agree with one another, that lends creedence to the assertion that the predictions are accurate. As the number of models increases, the level of agreement provides even more confidence in the predictive capabilities.

Other goodness-of-fit metrics exist, but in additon to being the ones used by Toal to develop rules of thumb, provide a good balance of assessment of trend and error. If the shape of the data aligns, the R^2 will increase. However, $RMSE$ will reduce the confidence level if there is a major discrepancy in magnitude between the two trendlines. This leads to Hypothesis 1.3.

Hypothesis 1.3 *For a sufficienctly large model set, as inter-model agreement increases, measured by R^2 and $RMSE$, confidence in the quality of the associated models, as represented by the probability of highest fidelity, increases.*

5.5.3 Calculation of Metrics and Normalization

The first step in the process of fidelity scoring by comparative data analysis is the calculation of R^2 and $RMSE$ for each pairwise model comparison. However, calculating regression metrics requires data that has aligned parameter values. The responses being used to calculate correlation and error metrics must be at the same design points or the metrics are not trustworthy. Fortunately, the models of this set are well-behaved, so all of the cases ran smoothly. However, there are times when additional effort is needed to be bring the data sets into alignment, using the following generic steps:

1. Given two models
2. For each variable, find shared range
3. Interpolate as necessary so each response value has a pair

As model set 2 if efficient and reliable, the data sets are completely aligned, so this will be discussed further in the next chapter.

In certain cases, such as with the I-beam model set, responses may already be developed for the same variable values and evaluation may proceed smoothly. This is, however, not the most realistic assumption in general. This will be adressed further in the next chapter.

Whether or not additional steps are required due to misalignment, once the pairwise correlation metrics are calculated, they can be normalized into fidelity scores between zero and one using the process shown below for a generic matrix of correlations **C**. Note that the diagonal is set to zero regardless of which correlation or error metric is used for consistency.

The sum of the rows of the absolute value of the transpose of **C** are saved as **R**. The sums are then added to the transpose of **C** to find **D**. **D** is then divided by the absolute value row sum of **D** minus **R** to account for the diagonals, normalizing the values by row. To find the score matrix, the diagonals are then subtracted to return the diagonal values to zero. The scores for each model are then the row values of the transpose of **D**, called **S**, not including the diagonals.

$$\mathbf{C} = \begin{bmatrix} 0 & R^2_{1,2} & \cdots & R^2_{1,n-1} & R^2_{1,n} \\ R^2_{2,1} & 0 & \cdots & R^2_{1,n-1} & R^2_{1,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ R^2_{n-1,1} & R^2_{n-1,2} & \cdots & 0 & R^2_{n,n} \\ R^2_{n,1} & R^2_{n,2} & \cdots & R^2_{n,n-1} & 0 \end{bmatrix}$$

$$\mathbf{R} = \sum_i |C_{ji}|$$

$$\mathbf{D} = \mathbf{C}^T + \mathbf{R}$$

$$\mathbf{D} = \frac{\mathbf{D}}{\sum_j |D_{ij}| - \mathbf{R}}$$

$$\mathbf{D} = \mathbf{D} - \text{diag}\,(\mathbf{D})$$

$$\mathbf{S} = \mathbf{D}^T$$

An important note is that for $RMSE$,

$$\mathbf{D} = -\mathbf{C}^T + \mathbf{R}$$

since RMSE improves as it gets smaller, where R^2 improves as it increases.

The correlation scores are normalized such that they are comparable to what was previous generated from expert opinions. As such, they can be used as a new set of samples using KDE to generate updated distributions, and by extension, model probabilities, that are now based on both the expert assessments and comparative data analysis.

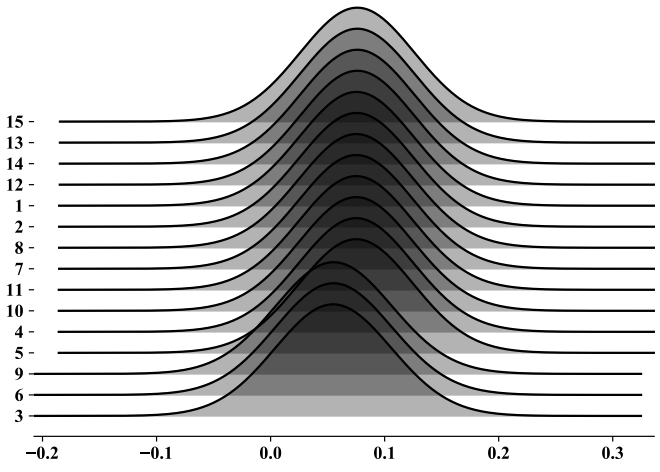
An additional consideration in adding these new rankings to the sample set is the weights. There will be $m - 1$ scores for each model in a set of size m, and two different metrics are used. Barring any additional information, the two metrics R^2 and $RMSE$ should have an equivalent total weighting. Additionally, the correlation scores should have an equivalent total weighting to the expert-provided scores. The weights are adjusted accordingly and an example of the resulting assessments are shown in figures for the linear buckling results of the initial model set.

The density estimates based on R^2 and $RMSE$ are shown in Figures 5.14 and 5.15. Notice that the order is different than it was from the purely descriptive assessment. Most noticeably for the R^2 scores, the Quad/Rod models stand out in their own group as worse than the rest of the models. This agrees with the previous observation that those model responses are nearly zero across the entire design space. The rest of the models are model difficult to visually distinguish since their medians are more similar. Additionally, the top four models in terms of R^2 are the same as from the descriptive assessment, leading to Observation 1.3.1.

Observation 1.3.1 *R^2 is an important metric for reliable multifidelity regression, as well as comparative data agreement. It also helps to lend creedence to Hypothesis 1.3, since the models believed to be the highest fidelity, having their own distinct grouping in the response, also score highest in terms of R^2.*

In addition, the order in terms of $RMSE$ is quite a bit different than the other orders. However, there is less variation in the median values for RMSE than for R^2. The 1-D models rank most highly in terms of $RMSE$ because, as can be seen in Figure 5.12, when the length of the beam is low, the magnitude of the 1-D models fall in the middle of the range. The shell and solid models ranked highest by the descriptive and R^2 scores are the lowest by $RMSE$. This leads to Observation 1.3.2.

Observation 1.3.2 *While Toal's rules of thumb focus on R^2, $RMSE$ must be included to properly adjust model fidelity assessment based on model agreement since, if the models completely agreed, the residuals would be low.*



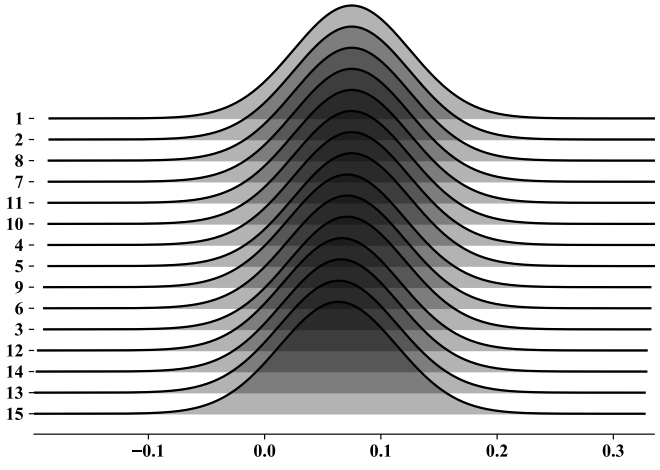
(a) KDE

Median	Model	ID
0.05514	Quad4/Rod	3
0.05549	Quad8/Rod	6
0.05555	QuadR/Rod	9
0.07500	Quad4/Beam	5
0.07500	Quad4/Bar	4
0.07500	QuadR/Bar	10
0.07500	QuadR/Beam	11
0.07509	Quad8/Bar	7
0.07509	Quad8/Beam	8
0.07532	Beam	2
0.07533	Bar	1
0.07598	Quad4	12
0.07599	QuadR	14
0.07600	Quad8	13
0.07602	Hex Solid	15

(b) Median Value

Figure 5.14: Medians and Distributions of R^2 Scores

Combining the two new sets of sets based on comparative data analysis, and joining this with the scores from descriptive assessment, the probability of being highest or lowest fidelity can be recalculated, as shown in Figure 5.16. Due to the agreement between most of the models, their probabilities become more similar. Additionally, the probability that



(a) KDE

Median	Model	ID
0.06350	Hex Solid	15
0.06386	Quad8	13
0.06558	QuadR	14
0.06650	Quad4	12
0.07095	Quad4/Rod	3
0.07095	Quad8/Rod	6
0.07095	QuadR/Rod	9
0.07485	Quad4/Beam	5
0.07485	Quad4/Bar	4
0.07486	QuadR/Bar	10
0.07486	QuadR/Beam	11
0.07490	Quad8/Bar	7
0.07490	Quad8/Beam	8
0.07491	Beam	2
0.07491	Bar	1

(b) Median Values

Figure 5.15: Medians and Distributions of $RMSE$ Scores

the Quad/Rod models (3, 6, and 9) are the lowest fidelity increase, despite most of the other values decreasing. Specifically for model 3, the probability of being lowest fidelity increases 64.4%, while the value for model 1 decreases 49.5%.

5.5.4 Initial Model Down-Selection

Looking further at Figure 5.16, the situation is more ambiguous than before adjustment. There are groups of models that have similar probabilities. One of the reasons for this can be ascertained by visual inspection of the response data in Figure 5.12. Many of the trends seems to be overlaying one another, confusing the assessments. Unfortunately, for problems with higher dimensionality, this might not be as easy to ascertain from a plot of the response data. However, the pairwise assessments of correlation and error can potentially provide an additional benefit by enabling an intial model screening.

While it is desirable for the R^2 to be high and the $RMSE$ to be low between multifidelity model selections, if they are in near-perfect agreement between two similar models,

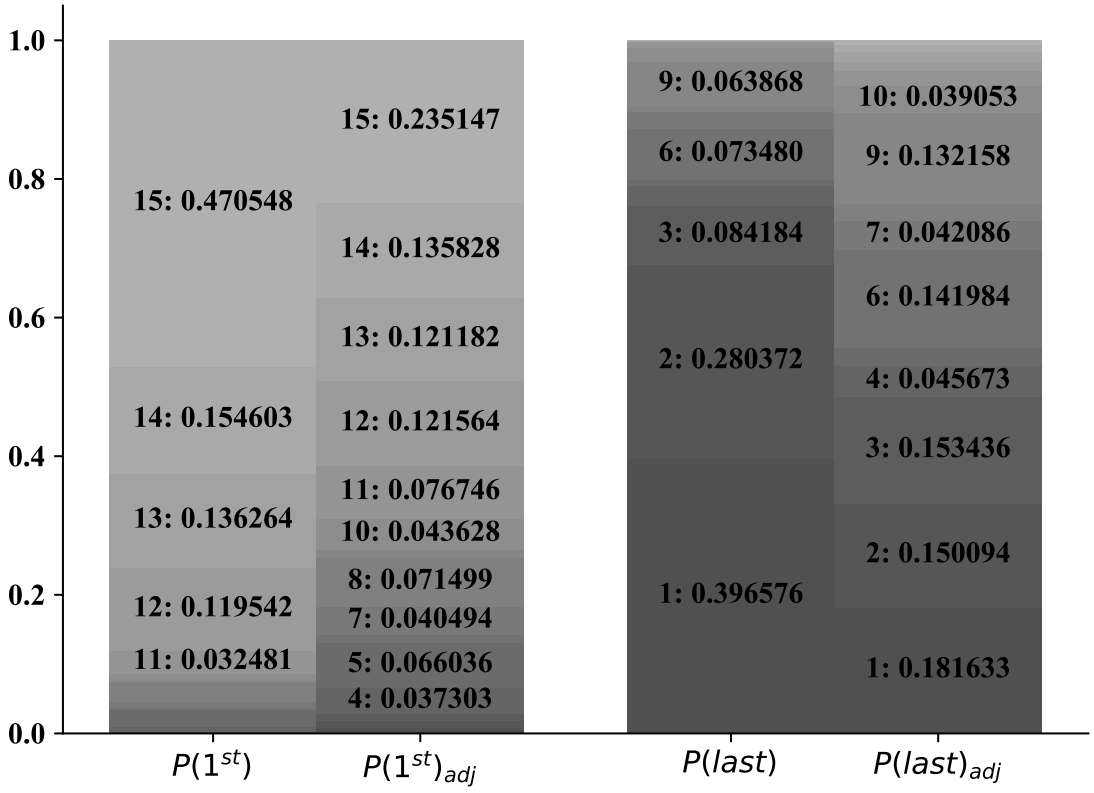


Figure 5.16: Initial Probabilities Versus Adjusted for Correlation and Error Scores

they are essentially duplicates. This could be used as justification for eliminating one of those models from consideration, especially when they are similar models that take a similar amount of time to run. The threshold of "very high agreement" is the subject of consideration and adjustment in each case. Toal recommended an R^2 higher than 0.9 for reliable multifidelity regression. However, an $R^2 > 0.98$ along with an $RMSE < 0.05$ could be evidence of an essentially duplicate model set, leading to the following observation:

Observation 1.3.3 *While other factors must be considered before removing any model from consideration, a very high R^2 and correspondingly low $RMSE$ between two similar models provides justification for removing one of the duplicate models from further consideration.*

The R^2 and $RMSE$ for fives subsets of models are shown in Table 5.9. In such a case, when the transverse web is defined with a shell element, the differences between

Models	Mean R^2	Mean $RMSE$
Bar, Beam	0.9999999	3.609e−5
Quad4/Bar, Quad4/Beam	0.9999999	7.389e−6
Quad8/Bar, Quad8/Beam	0.9999999	1.055e−3
QuadR/Bar, QuadR/Beam	1.0	2.151e−20
Quad4/Rod, Quad8/Rod, QuadR/Rod	0.9999999	7.189e−6

Table 5.9: Model Set 2 Duplicate Models

the flanges being represented with bar elements is indistinguishable from the use of beam elements. Additionally, the three quadrilateral web, rod cap, models produce indistinguishable results, so the Quad4/Rod and Quad8/Rod models can be removed. Similarly, the two one-dimensional representations are similar enough to warrant removal of the Bar model type. These groupings can also be seen in the median values in Figures 5.14 and 5.15. This allows for the removal of 6 models from consideration, leaving a set of 9 models. As the model set is down-selected, the models should be appropriately excluded from the descriptive fidelity orders in Table 5.8.

If there was a doubt as to whether or not a model should be removed, the decision can always be delayed until a more thorough decision-making process is undertaken. A contributing factor of this is implied by the three response scenarios generated for this model set. A true model decision-making process can be multi-attribute, so a model should not be completely removed from consideration without taking into account all of the potentially relevant solutions. This will be discussed further in coming sections.

Quad/Rod Models

The model probabilities could then be re-calculated for the down-selected set of 9 models based on the experts assessments and comparative data analysis. However, as well as identifying duplicate models, this method helps to identify models who disagree with the others to the extent that further investigation is required. Either the model does not adequately represent the correct behavior to be appropriate for a given problem or the model

just requires troubleshooting and debugging. In this case, the clear outliers in terms of R^2 are the Quad/Rod models.

Two of the Quad/Rod models were removed as duplicates, so that leaves QuadR/Rod as the only one remaining of that category. Looking at the linear buckling responses in Figure 5.12, all of the responses for this model type are nearly zero, on a different order of magnitude from the rest of the models. What makes this interesting is that essentially the same model agrees with the majority of the set for the other two problem definitions, linear static deflection and normal modes.

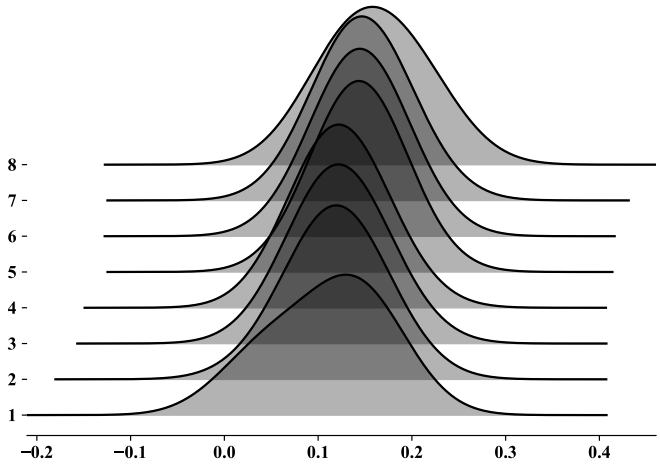
This model was specifically developed to represent the problem that can occur in the common method of model decision-making: simply taking a previously used or partially developed model and applying it to a new problem. These Quad/Rod models are capable of representing the appropriate behavior. In fact, it is given as one of the example models for linear buckling analysis in the Nastran Linear User's Guide[99]. The issue is that Rod elements are designed purely for axial and torsional loads, meaning they cannot carry a load in the transverse direction. As such, the buckling limit that is being reported is the buckling limit of the shell web elements, which is essentially zero.

Finding the buckling limit looks for the path of least resistance, so along the direction of the web, the Quad and Rod elements can resist the load, but to either side, the flanges are what resist buckling. Since the Rod elements cannot carry load in that direction, they do not correct represent the flanges. Bar and Beam elements are designed to carry a transverse load, which is why those models behaved correctly from the start.

As described in the Linear Users's Guide, a Quad/Rod model can provide a good estimation of the linear buckling limit, but additional boundary conditions are required to prevent side-to-side motion. However, these boundary conditions are not required for the other problem definitions or the other element types, so it would be easy for someone to use a previously developed model and overlook such issues. This leads to the Observation 1.3.3.

Observation 1.3.4 *Using correlation and error metrics to assess the comparative fidelity of models in a set can help to identify models that need troubleshooting, debugging, or do not represent the appropriate phenomenology.*

This behavior is easy to find simply by looking at the results in this case, since there is only one variable. Importantly though, if the dimensionality were to increase, it would not be as obvious when a model is behaving differently from the rest. Therein lies one of the powers of these methods, since R^2 and $RMSE$ could still find those differences even with increased dimensionality. If the boundary conditions were to be added to the Quad/Rod models, their results would be in line with the other 1/2-D hybrid topology models. As such, for simplicity, the QuadR/Rod model should be removed from the set, leaving only eight fidelity levels for consideration, described in Figure 5.17.



(a) KDE

Median	Model	ID
0.1132	Beam	1
0.1189	Quad4/Beam	2
0.1223	Quad8/Beam	3
0.1231	QuadR/Beam	4
0.1437	Quad4	5
0.1445	Quad8	6
0.1475	QuadR	7
0.1606	Hex	8

(b) Median Values

Figure 5.17: Correlation-Adjusted Fidelity Estimates for 8-Model Set, Sorted By Median

The results of the down-selected model set are shown in Figure 5.18. The comparison of model probabilities are shown in Figure 5.19.

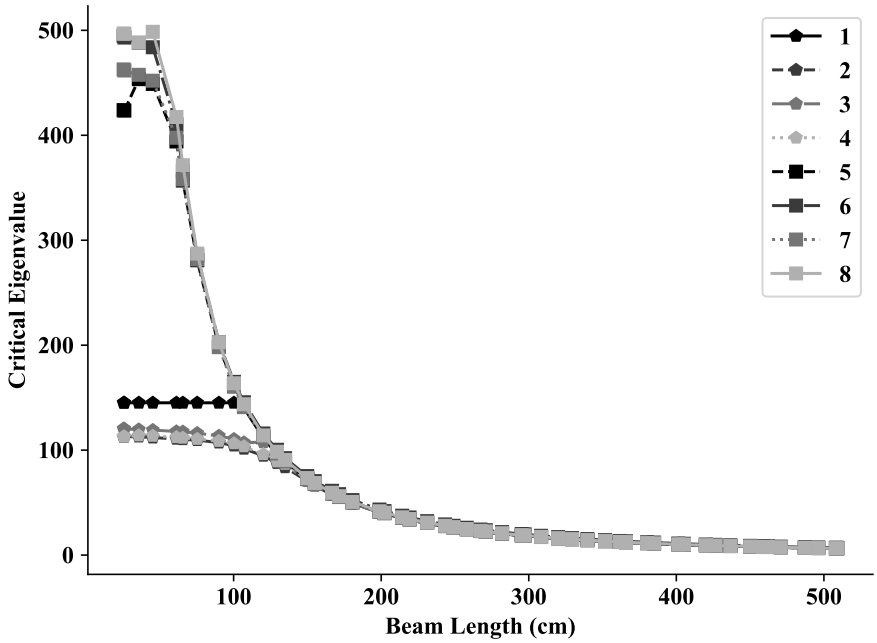


Figure 5.18: 8-Model Linear Buckling Results

5.5.5 Comments on Correlation Scoring

In this case, it is commonly agreed upon that the solid element type should provide the highest fidelity, most generally accurate, model. However, after including the correlation and error scoring metrics, while the Hex model still has the highest probability of being the highest fidelity model, it is less obvious than just from the descriptive assessment. This is due in part to the way the metrics work, specifically the $RMSE$.

However, this adjusted assessment is specific to the linear buckling problem definition. The descriptive assessment should account for the problem definition, but there is only so much tailoring that can be done by experts prior to the infusion of quantitative results. The process of providing a descriptive fidelity assessment is analogous to defining an informative prior distribution for model fidelity in that the relative ordering of the models should be as correct as the qualitative assessment allows. This is why it is important that the metrics be linguistically specific and the process be straightforward, it allows for a better initial understanding of the model set than would otherwise be possible. Once model data becomes

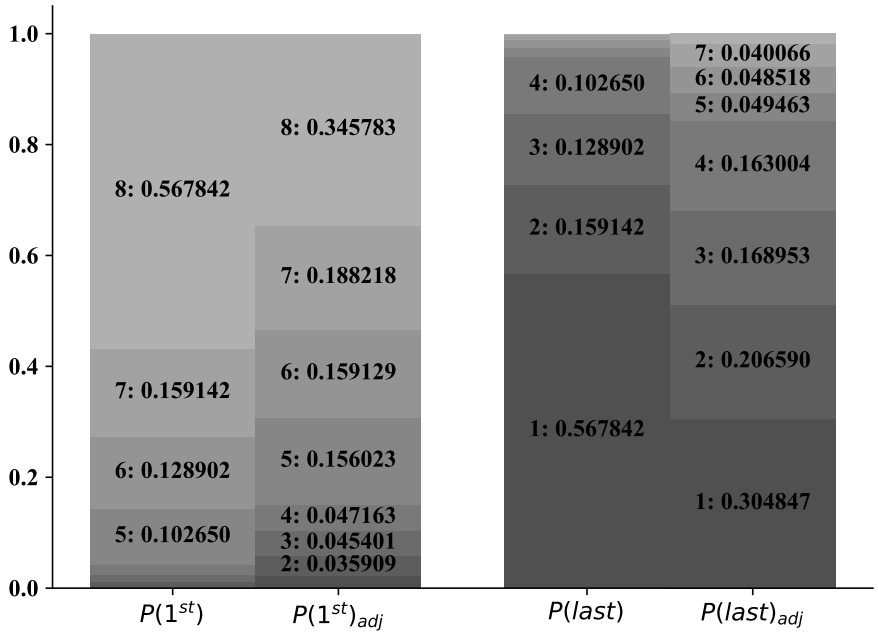


Figure 5.19: 8 Model Probability Comparison

available, the comparative data analysis is used to find an analogous posterior distribution. The relative magnitudes of the fidelity estimates are adjusted in a way that accounts for how much the models agree or disagree with one another.

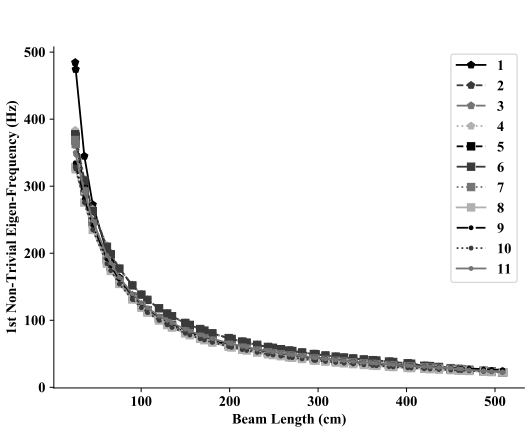
Linear Static

Applying the comparative data analysis methods to the linear static problem, duplicate checking is crucial. Given the R^2 and $RMSE$ scores, 12 of the 15 models appear as duplicates. Interestingly, if all 12 are removed, the hybrid topology approach is removed entirely, as their results overlay the shell model results. Even after removing the models that meet the default criteria for duplicates ($R^2 > 0.9$ and $RMSE < 0.05$), the three remaining models still have very similar results that could be considered as duplicates if the tolerances were adjusted. The remaining models have a mean R^2 of 0.995 with a minimum of 0.99, and a mean $RMSE$ of 0.215 with a maximum of 0.417. As such, regardless of how many models are left in after checking for duplicates, the correlation scores are going to adjust

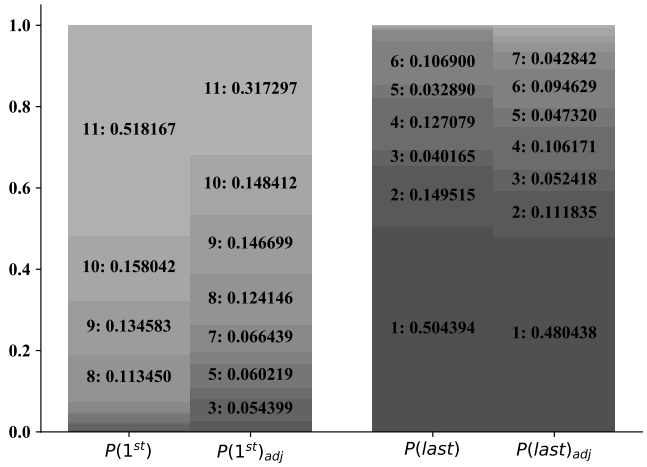
the fidelity distributions and associated probabilities toward one another.

Normal Modes

The results for the normal modes problem are by and large not that different from the linear static case, with one major exception. From Figure 5.13, it is obvious that the Bar model diverges from the rest as the length decreases. This is presumably due to the change in the length/depth ratio discussed in the model set definition, since the assumptions of the bar model break down when the beam is no longer beam-like.



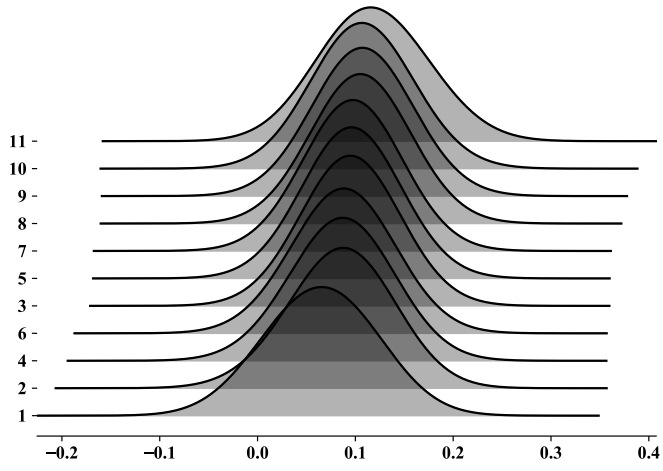
(a) Normal Modes



(b) Fidelity Probabilities

Figure 5.20: Normal Modes Results and Fidelity Probabilities for 11-Model Set

After removing the Bar model from the set, the Quad/Rod models again stand out as different from the majority of the remaining models, since they can resonate side-to-side with less impediment than the other models. The vibratory frequency, however, is not as obviously different upon initial inspection as in the linear buckling case. Removing the Quad/Rod models leaves the results shown in Figure 5.20a and the adjusted distributions shown in Figure 5.21. In this case, the adjusted distributions and associated probabilities, shown in Figure 5.20b, are not that different from those from the descriptive assessment. In fact, the only significant difference is the reduction in the probability that the Hex model



(a) KDE

Median	Model	ID
0.0646	Beam	1
0.0858	Quad4/Bar	2
0.0862	Quad8/Bar	4
0.0876	QuadR/Bar	6
0.0948	Quad4/Beam	3
0.0960	Quad8/Beam	5
0.0973	QuadR/Beam	7
0.1052	Quad4	8
0.1074	Quad8	9
0.1074	QuadR	10
0.1180	Hex Solid	11

(b) Median Values

Figure 5.21: Normal Modes Adjusted Fidelity Estimates for 11-Model Set

has the highest fidelity, since the results agree with the other models. Examining how comparative data analysis enables duplicate removal, initial down-selection, and fidelity adjustment for these three problem sets leads to Conclusion 1.3.

Conclusion 1.3 *Correlation and error metric fidelity scores increase the understanding of the model set, and, when combined with expert assessments, can provide better model fidelity probabilities than would otherwise be available.*

5.6 Multifidelity Model Rankings

5.6.1 Research Question 2

Model fidelity density estimates and probabilities provide an understanding of where each model fits into the set. However, it does less to enable the informed selection of a multi-model combination. This leads to the following research question:

Research Question 2 *How can model fidelity evaluation be extended to rank ordered multifidelity model combinations for model selection?*

5.6.2 Observations and Research Question 2.1

Model subset selection must be in a particular order. As evidenced by the process for eliciting expert opinion, there is an ordering of the relative fidelity and even of the relative accuracy of the models. Additionally, a particular order is needed for practical purposes, such as in the development of a Co-Kriging regression. This complicates the process as the permutations of the model set must be evaluated instead of simply the combinations. Based on this, the following research question is developed:

Research Question 2.1 *Given a set of model probabilities, how can ordered combinations of models be ranked in terms of fidelity*

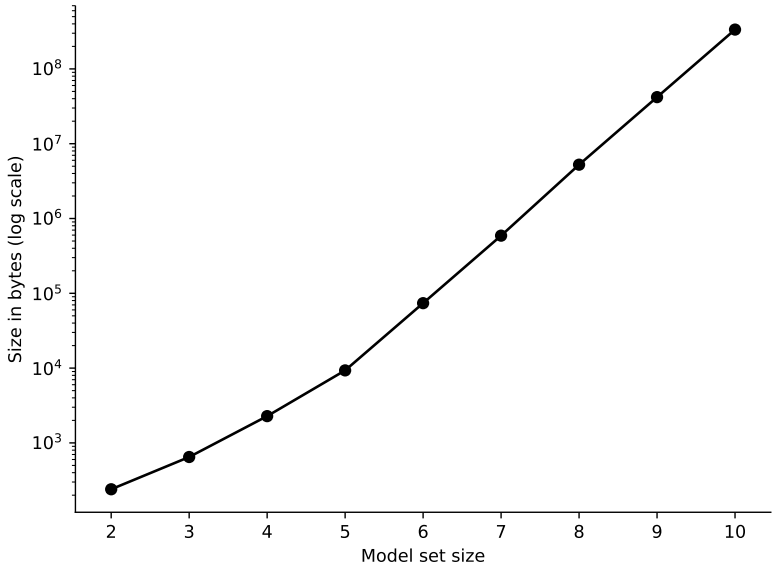
An example of the difficulty of ordered sets is evident even for the down-selected model set. Given eight models, there are $109,600$ permutations, but only 247 combinations. Even if the evaluation or each permutation is simple, the amount of time required to evaluate the full set and the memory requirements for storing the results grows at an essentially exponential rate with the size of the model set. The necessary computational resource required is prohibitive.

This is shown in Table 5.10. Evaluated in Python, for each model set size, all of the possible permutations are given a score of 1.0 and stored in a dictionary. The keys are a tuple of integers denoting the order, e.g. *(1, 2, 3)*, and the values are all 1.0, as previously stated. The amount of RAM required for each is plotted on a semi-log scale, and appears nearly linear, suggesting a nearly exponential relationship.

While 320 megabytes is typically much less than the amount of RAM in a typical modern desktop, at that rate the memory requirement should be over one gigabyte at around thirteen models. Additionally, and importantly, the time required to iterate through every permutation scales similarly to the memory requirement. For a size above ten, even this evaluation where every value is simply set to 1.0 goes from a fraction of a second to minutes. This would only be worse for a complicated scoring process, and, to account for this,

Table 5.10: Memory Required to Retain Multifidelity Order Scores

Set Size	Memory
2	240.0 B
3	648.0 B
4	2.23 KB
5	9.1 KB
6	72.09 KB
7	576.1 KB
8	5.0 MB
9	40.0 MB
10	320.0 MB



a soft limit is put in place.

When the size of the model set is over 9, permutations are only assessed up to a length of 5. This is fairly easy to justify as a down-selection above a certain size has serious implications for the development of the selected models and keeping track of all of their associated data. Additionally, this logical extension leads to Requirement 2.1.1.

Requirement 2.1.1 *Multifidelity model order rankings, whether inherently or by extension, should account for the difficulty of developing and maintaining large model sets in parallel.*

Multifidelity methods are primarily used for two reasons, efficiency and robustness. In either case, the multi-model combination is not assumed to be higher fidelity than the best model in the set. In fact, some fidelity is always traded in multifidelity combination, reducing coverage of the design space for one model in order to boost the efficiency or get a constrasting opinion. Additionally, using multifidelity models to increase robustness is primarily only relevant for stochastic models. This leads to Requirement 2.1.2.

Requirement 2.1.2 *The score for a multifidelity collection should be based on a combination of the scores of the individual models, and whether they were selected in the correct order, from highest to lowest fidelity.*

5.6.3 Hypothesis 2.1

A ranking of a particular order implies the process of describing the quality of a particular series of decisions. Given m models and expecting an order of size n, n selections must be made without replacement from the avilable options. The model probabilities developed in the previous sections represent the chance that a model is the best choice from a set. Extending this, the model probabilities can be re-calculated for each available combination of available sets. As mentioned above, this is a sizeable number of calculations, but not nearly as many as the number of permutations.

This leads to the following hypothesis:

Hypothesis 2.1 *For an ordered combination from a model set, if the probabilities that each model is the highest fidelity available are combined, then the quality of the ordered down-selection can be scored and compared to the other permutations.*

5.6.4 Multifidelity Scoring From Individual Probabilities

For the notional model set, $m = 4$, the number of combinations of size one to m is fifteen, and the number of permutations from size one to m is a very manageable 68. If the down-selection, ordered from lowest to highest fidelity, is

$$\text{Model 1} \rightarrow \text{Model 2} \rightarrow \text{Model 3} \rightarrow \text{Model 4}$$

then the scoring value of each selection can be defined as follows:

- $P(\text{Model 1} \in [1, 2, 3, 4])$
- $P(\text{Model 2} \in [2, 3, 4])$

- $P(\text{Model } 3 \in [3, 4])$
- $P(\text{Model } 4 \in [4])$

Note that the last probability when size of permutation $n = m$ is always going to be 1.0 since there is only one option left. As this is a series of independent decisions, the combined probability that all of them occur is the product of the probabilities, as described below. Using the product of the probabilities as the scoring of the permutation has a number of benefits. For one, all that is needed is the current estimates of fidelity probabilities for each combination. Additionally, as the size of the permutation grows, it becomes less likely that the quality of the selection is improving. Put another way, it is more likely that one of the included models is a good choice. As such, the product of a large number of probabilities is unlikely to be highly ranked, which directly addresses Requirement 2.1.1; The scoring system is inherently dependent on the size of the permutation.

However, directly using the model probabilities does not adequately represent requirement 2.1.2. If the two highest models have probabilities of being the highest fidelity of 0.3 and 0.29, then their combined score will be 0.087, much lower than either of the single-model values. To determine what modifications need to be made, the notional model set can be used.

5.6.5 Multifidelity Scoring of Notional Model Set

For the notional 4-model set, there are 64 total permutations of length 1 through 4. The small size of options means that the full set can always be evaluated. The probabilities of being the highest fidelity model when all models are available for selection is what was shown in the first columns in Figure 5.5. As mentioned above, the probability of selecting a model when only one is available is always 1.0, so the model probabilities must be re-evaluated for the eleven combinations of size two to four.

Given all of the model probabilities and a particular permutation, assuming that the permutation is ordered from highest to lowest, the multifidelity score is calculated by mul-

tiplying the appropriate series of probabilities, as described in the previous section. An example follows:

For ordered set: (4, 3, 2, 1)

$$\begin{aligned}\text{Score} =& P(\text{Model 1} \in [1,2,3,4]) \times P(\text{Model 2} \in [2,3,4]) \\ & \times P(\text{Model 3} \in [3,4]) \times P(\text{Model 4} \in [4]) \\ =& P(M_1 > M_2 \cap M_3 \cap M_4) \times P(M_2 > M_3 \cap M_4) \times P(M_3 > M_4) \times 1.0 \\ =& 0.36396 \times 0.45298 \times 0.54556 \times 1.0 \\ =& 0.089947\end{aligned}$$

Or for 4 and 3:

$$\text{Score} = 0.369396 \times 0.45298 = 0.167329$$

Even when selecting the two highest available fidelity models, the combined score is much lower than either of the single-model scores. To account for this, the each set of probability of best available could be normalized to the highest value. In that case, the previous example, selecting the highest available each time, 4, 3, 2, 1, would result in a combined score of 1.0. This is also not preferable, since the size of the aggregated group no longer has an effect on the final score.

If, instead, the probability of best available is normalized to the highest value, then multiplied by the probability of highest fidelity for the full array of models, normalized to the highest value, the size of the combination is taken into account, and the original ranking is taken into account, as follows:

For ordered set: (4, 3)

For full set:$P(1^{st}) = 4 : 0.364, 3 : 0.303, 2 : 0.182, 1 : 0.151$

Normalized to 0.364: $= 1.0, 0.832, 0.500, 0.416$

Score $=1.0 \times (1.0 \times 0.832) = 0.832$

One of the benefits of normalizing the scores is that the single highest fidelity model will always score a 1.0. This provides a baseline for comparison of all other combinations. Normalizing and then scaling means that the combination of two models will result in a score not drastically different from the single-model scores, but also dependent on whether the models were selected from highest to lowest fidelity without skipping any models.

However, there is still an issue when selecting the two highest fidelity models. In the case shown above, the combination 4, 3 scores 0.832, but 3, 4 would also score 0.832. Additionally, this is exactly the same score as model 3 by itself. This is a special case that only pertains to the top two models in the set, but is still in disagreement with the requirements. To account for this, the normalized probability of best available should be scaled by the average of the score and the highest possible score of 1. This means that the combination of models 4 and 3 will score $1.0 \times \frac{1+0.832}{2} = 0.916$, whereas 3 then 4 will score $0.832 \times 1 = 0.832$, so order is preserved. Additionally, for combinations involving all of the models, or length m, the last selection will no longer score a 1, so orders of length m will not score always score identically to orders of length $m - 1$.

This method of calculating a fidelity score for single and multi-model ordered combinations meets all of the requirements and is described in full detail in Algorithm 2.

To test out the fidelity scoring method, the four notional cases previously described as used: all increasing, no scope, fixed scope, and reversed scope, or realistic. The fidelity

Algorithm 2 Fidelity Score for Multi-Model Ordered Combination

Require: Pairwise $P(X > Y)$ for model set of size m {Get single-model scores}
  $P_{(n=m)} \leftarrow P_{1^{st}}(\forall models)$
  $P_{(n=m)} \leftarrow P_{(n=m)}/max(P_{(n=m)})$ {Normalize to highest value}
  {Get non-ordered multi-model values}
  for $i = 1$ to $m - 1$ **do**
    for all $combination$ in $combinations(models, size = i)$ **do**
      $P_{combination} \leftarrow P_{(1^{st} \text{ available})}(combination)$
      $P_{comb} \leftarrow P_{comb}/max(P_{comb})$ {Normalize}
      $w = P_{(n=m)}$(highest fidelity available)
      $P_{comb} \leftarrow P_{comb} \times \frac{1+w}{2}$ {Scale value}
    end for
  end for{Get (order, score) pairs}
  for $i = 2$ to m **do**
    for all $order$ in $permutations(models, size = i)$ **do**
      $available = sorted(order)$
      $score = 1$
      for all ID in $order$ **do**
        $score \leftarrow score \times P_{(1^{st} \text{ avail})}(ID)$
        Remove ID from $available$
      end for
      $scores[order] \leftarrow score$
    end for
  end for
  return scores

probabilities with resepct to the full set for these four cases were shown in Figure 5.5. Figure 5.22 shows the single model scores and the highest multifidelity scores for each case.

For all of the cases, as defined previously, the single most highly ranked model is always given a score of 1.0. The other single models are given scores based on their probability of highest fidelity, normalized by the highest model's probability value. As expected, the combination of the two highest scoring models in the correct order shows a loss in fidelity from the highest option, but is better than the second highest option. However, order is taken into account, since choosing model 3 before model 4 limits the capability of the overall combination by the starting score of model 3, both of which address Requirement 1.1.2. Additionally, as previously discussed, larger combinations are going to inherently

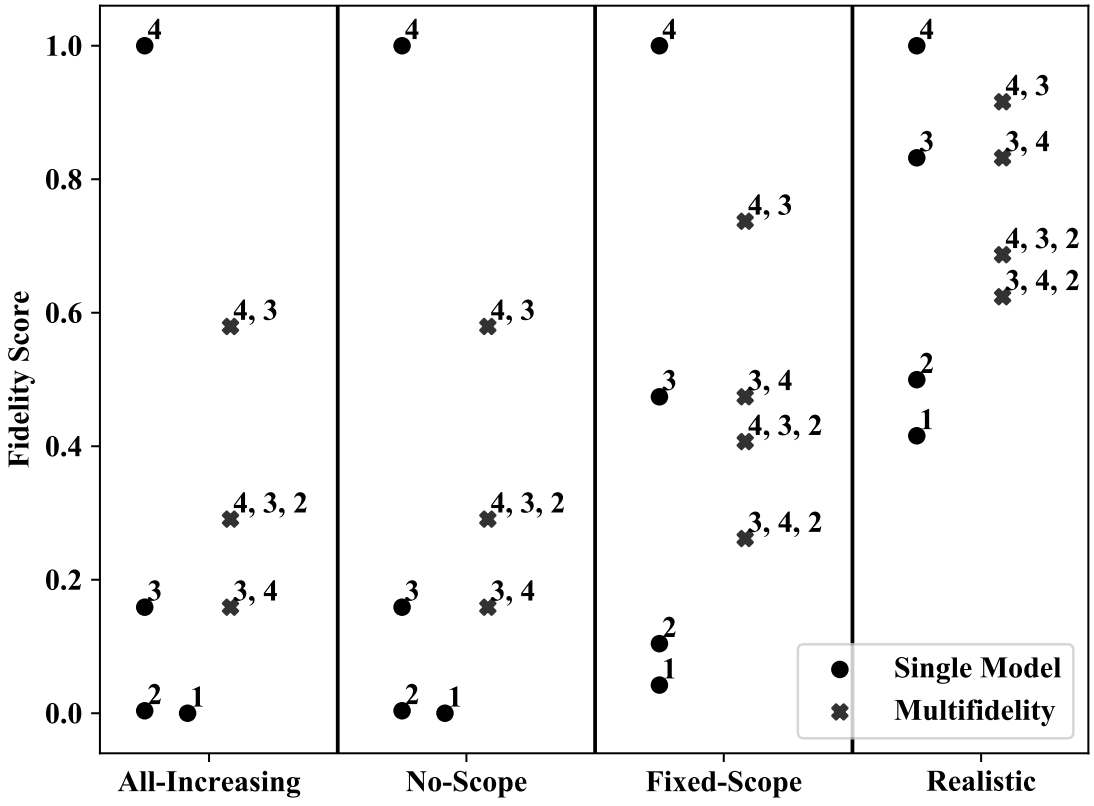


Figure 5.22: Fidelity Scores for Four Cases of the Notional Model Set

generate a lower score since the at least one of the models in the order is more likely to be individually low-scoring, addressing Requirement 1.1.1.

This shows that multifidelity scores can be generated for any given model ordering based only on expert assessment. If more data, such as correlation-based scores, are included, the analysis can be performed again. The adjusted scores can be used on their own for decision-making, or some information may potentially be gleaned from comparison of the original assessment to the adjustment scores, as with the individual model probabilities.

These scores lead to an increased understanding of the model set, even based purely on the initial descriptive assessment, and enable a model down-selection process. However, a true model selection process needs to incorporate the cost of the given selection, which will be discussed further in a later section. For the notional model set, all of the requirements for fidelity scoring are met, but the methods will continue to be tested using the results of

model set 2.

5.6.6 Multifidelity Scoring of I-Beam Model Set

Applying the multifidelity scoring methods to model set is similar, but more intensive. For the original size of fifteen models, there are $32,752$ combinations and $3,554,627,472,090$ permutations. As mentioned before, not all of those permutations would be evaluated for efficiency, but the evaluation time would still not be negligible. Fortunately, after initial down-selection, only eight models remain in the set, leading to 247 combinations and $109,608$ permutations. Compared to the 64 permutations of the notional set, this is still a large enough number to not be trivial, but is much more feasible than scoring all 3 trillion options.

Generating the probability of highest fidelity for all of the possible combinations and multifidelity scores for all of the permutations, for a set of this size, on a normal desktop, takes a matter of seconds. First, the orders for model set two are scored based on the descriptive fidelity assessment, and the single model scores as well as the top 20 multifidelity scores are shown in Figure 5.23.

Because models 5-7 are grouped together, multifidelity combinations including those three models score relatively highly. The 2-model combinations score the most highly, addressing Requirement 1.1.1. Again, as with the notional case, the correctly ordered combination of the top two highest scoring models (8, 7) falls in between the two single-model scores, while the reverse-ordered combination (7, 8) does not improve on model 7's fidelity score, addressing Requirement 1.1.2.

Referring back to the probabilies shown in Figure 5.19, the initial fidelity assessment gives model 8, the Hex model, over 50% probability of being the highest fidelity, and the shell models (models 5-7) are between 10 and 16%, with everything else much lower. As such, the top 20 multifidelity orders based on that assessment only contain those four models. Based only on the descriptive assessment assessment, the decision is really how

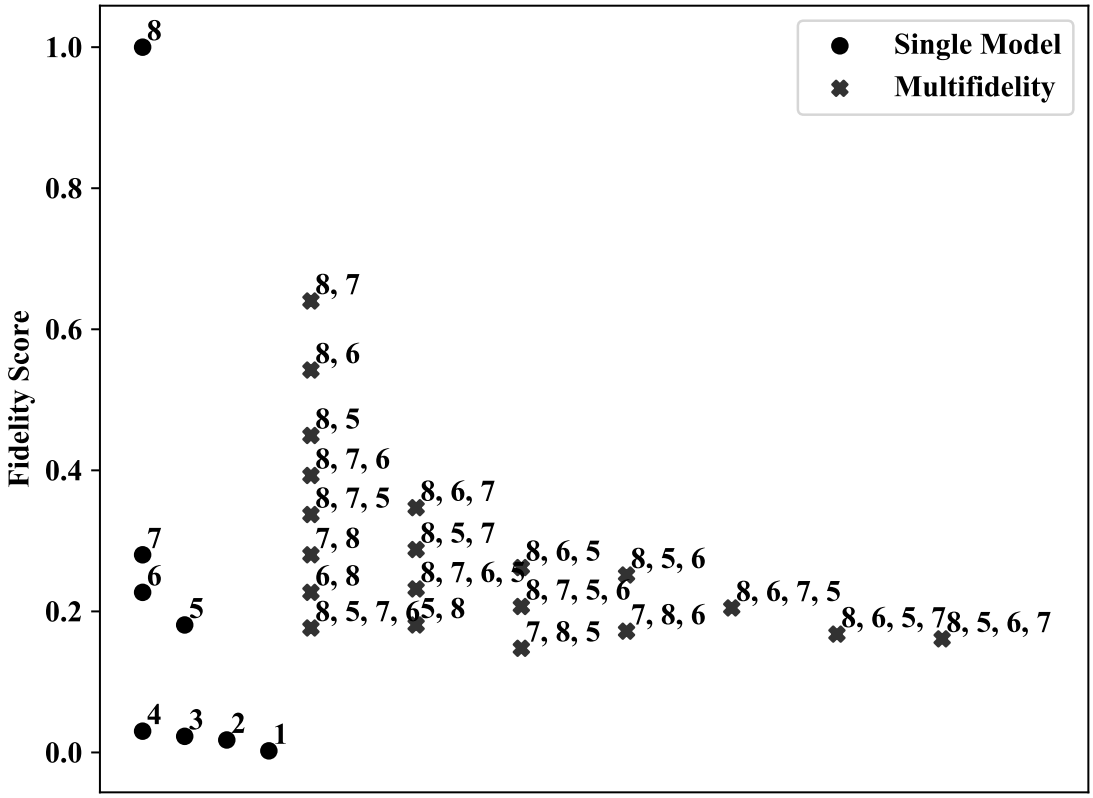


Figure 5.23: Descriptive Fidelity Scores for Down-Selected Model Set 2

many models the user can afford to run and keep track of, but most any combination of the solid model, potentially with a shell model, can be justified.

Taking the comparative data assessment with respect to the linear buckling results into account, the adjusted single and top 20 multifidelity scores are shown in Figure 5.24. Since some of the models are similar to one another, the probability of highest fidelity is increased from the descriptive assessment for models 2-7, however, they are still grouped together as before. The multifidelity orders that score highly are a similar list to those from the descriptive assesment, but the scores have changed based on the change in single model scores. Since models 5-7 now score closer to model 8, the 2-model combinations are more simlar.

As with the notional model set, model set 2 helps to prove out that the fidelity scoring method meets the prescribed requirements and scores single and multifidelity combinations

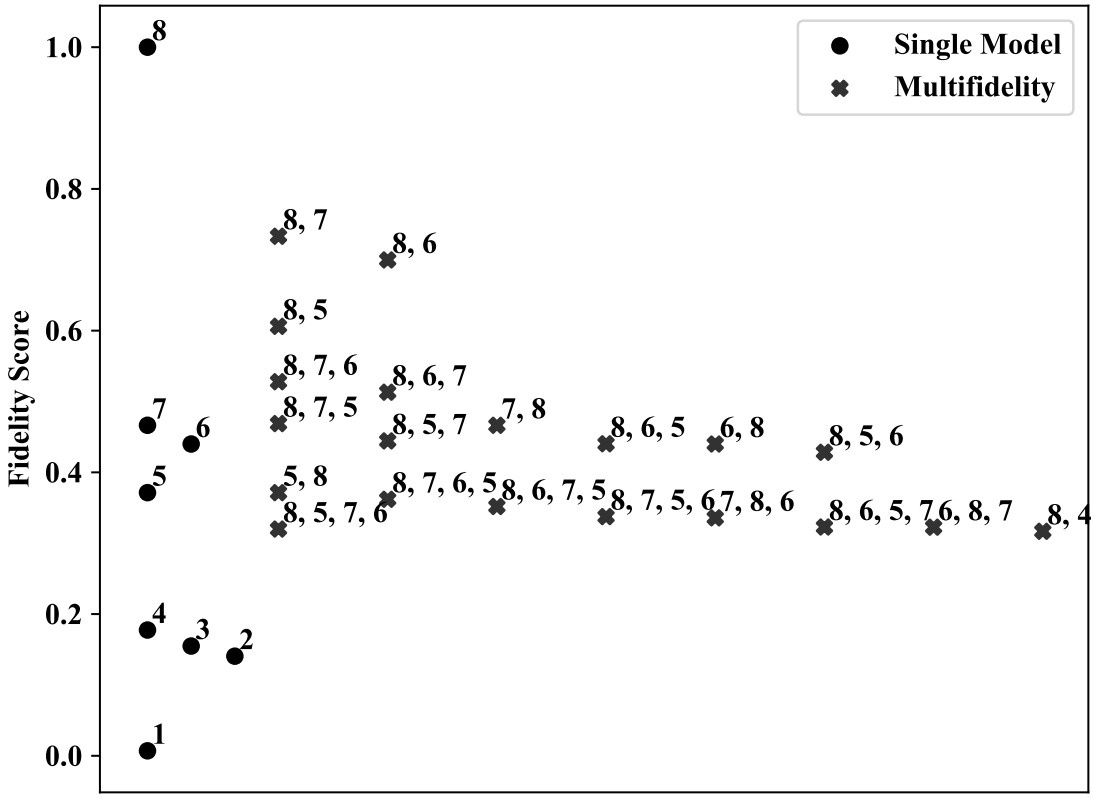


Figure 5.24: Descriptive Fidelity Scores for Down-Selected Model Set 2

in a way that advances understanding of the model set for decision-making. The score for a combination of models is dependent on the probability that model is the best available model, having removed previously selected models from consideration. This helps the score to represent whether the models were selected in the best possible order.

Additionally, the score is dependent on the size of the ordered combination, since as the size of the order increases, there are likely to be less quality options remaining. In addition, the normalization method implemented makes it such that the combination of models remains simlar in score to the single-model options, instead of much lower as they would be in the un-normalized case. However, aiming for the highest fidelity is not the only goal in model down-selection, as will be addressed in the next section.

5.7 Required Effort in Model Selection

As mentioned before, the efficiency of a model plays heavily in the model selection process. If one model takes much longer than another to generate, execute, or process, then that would make it less desirable based on the time schedule of the project. If evaluation of even a handful of cases expends the project's entire computational budget, then the fidelity of that model becomes less important. Minimal insight could be gleaned from those few cases, and the user would be better off having more cases from a lower fidelity but higher efficiency model. If a single model of the desired fidelity is not efficient enough, the user must be able to compare the loss of fidelity with the improvement in efficiency due to moving to a lower fidelity model or multifidelity combination.

In general, this leads to a need for a multi-attribute decision making (MADM) process, which this work seeks to enable. To reiterate a previous point, fidelity can even be a multi-attribute case since it is scenario, or response, driven. If multiple responses are important, the fidelity of each should be assessed, and a multi-attribute decision-making process could be undertaken simply for the various estimates of fidelity.

There are many MADM approaches, however, to enable their use, the single and multi-model ordered combinations must be scored in terms of cost or efficiency for comparison to model fidelity scores, leading to Research Question 2.2.

Research Question 2.2 *How can the required model cost be scored to allow for comparison to estimated model fidelity in a decision-making frramework?*

5.7.1 Measures of required effort

Model Degrees of Freedom

Before a model is to a phase of development where it can be run and the duration monitored, there is an additional way that the required effort can be measured: degrees of freedom. Typically this term comes up in discretized numerical modeling with respect to the mesh

and its constraints, but it also has a different meaning. Low fidelity, early conceptual design models may only use a handful of parameters to define one aspect of the system.

For example, an aircraft wing could be defined using simple planform parameters such as chord, span, sweep angle, taper, etc. However, as the resolution increases, many more parameters are required to define the exact geometry of the outer mold line (OML), layout of the ribs and spars, stiffener design across the skins, spars, and ribs, and so on[106]. Additionally, for a high abstraction model, the material may be broadly defined using the elastic modulus, density, Poisson's ratio, and a stress limit. However, as the abstraction is decreased, directional stress limits, fracture limits, and other allowables, knockdowns, etc. may be needed to understand the behavior of the structure[107].

This is brought up to point out how, under different circumstances, trying to bring a high fidelity model forward may put you in the position of needing information that is not yet available, or requires answers to decisions that have not yet been made. Because of this, variables are set to defaults, which can add uncertainty that contradicts the reason why the model was brought forward in the first place. If the minimum required set of variables can be enumerated, they could be treated as a surrogate for runtime prior to the evaluation of the first case. However, this will not be used further in this work.

Analysis Cost

The required effort for using a certain model is dependent on a number of factors. Typically what is thought of is the actual analysis time required by the numerical solver or solvers. This is for good reason, since when runtime is viewed as an impediment, it is because the user has to wait on the solver for an excessive amount of time before the result can be viewed. However, analysis time is not the only factor.

Generation Cost

Often, the amount of time required to create or generate a model is also burdensome. It is sometimes not viewed as such when it is a more interactive process, so the time is not simply spent waiting. For many finite element structural or computational fluid dynamic models, setting up the input file can take days or even weeks as a manual process. Especially for manual processes, the exact amount of development time required is difficult to measure and even more difficult to estimate. This is another factor as to why the model generation part of the required cost is overlooked.

Fortunately, when a process needs to be repeated or parameterized, my steps have been made to improve the capability of process automation, as evidenced by the RADE architecture used to develop the aircraft model set. Not only does automation mean that the setup process can be more reliable and faster, it can also be setup to run when a user is not available, such as overnight. This is especially beneficial as, for example, even as an automated process, mesh generation can be a lengthy process for some CFD models.

Post-Processing Cost

Another attribute that has to be considered is the post-processing required for a given model. Depending on the complexity of the response and the capability of the software, this can be simple or very difficult.

If the response of interest is something that the software provides automatically through the API, post-processing is easily automated. However, there are times when the response is easily assessed visually, but setting up an automated process to gather the same data requires its own set of research. This is brought up to emphasize that post-processing time can also be an important contributor to the cost element of model selection. Additionally, many of the same issues of developing manual or automated processes also apply to post-processing.

There are additional complications regarding the automation decisions made in devel-

opment of models to be included in a model set as described in this work. Depending on the complexity of the model and the availability of previous frameworks to build off of, automation can be a difficult process, as mentioned above. If only a small number of cases need to be run, it may not be worth it to automate the entire process. However, if a model is to be selected for use in a multifidelity design process, automation would be preferred or maybe even required. This adds to the difficulty of generating data for use in a decision-making process, which reiterates why the processes in this work to estimate fidelity and enable decision-making need to be streamlined and make efficient use of whatever data is available.

5.7.2 Duration Processing

The different measures of required effort typically come as durations of time. Situation-specific requirements such as available computational power and software licenses need to be considered, but are not as widely applicable. Saving or estimating the time required to generate, analyze, and process a model is not something that is always automatically done, but should become second nature for model developers. It is important for troubleshooting, providing an additional level of understanding of a model, but also in proving that the model is worth continued development.

Gathering required durations for a model is fundamentally different than saving model responses. To begin with, they are experimental data points, and there will always be some level of noise. If the computer is being taxed by some unrelated process, then the time returned can be drastically altered in a way that is not representative of the model, meaning that outliers need to be removed before processing. The complications with gathering reliable computational durations is lessened in certain cases. For example, when using Nastran, the log file returns some information regarding the difference between wall time, or the amount of time that passed on a clock, and CPU time, the amount of time spent purely to process the solution. When available, the CPU time should be used, as the aleatory

uncertainty caused by extraneous processes has less of an impact on this value.

Additionally, more can generally be said by one runtime value than by the response at one design point. Additional runs can help to filter out the effect of noise and determine if model runtime is dependent on the location in the design space, but even a single run can provide a cost estimate for a model. This lessens the burden of data avilability, as the model need only be at the level of development required to run a small number of times.

While the three sets of durations should all be recorded, they may not all be crucial to the model selection process. It is not unlikely that all of the models in a set may be generated or post-processed in a similar way, so their values would not show a large amount of discrepancy. Conversely, but less likely, all of the models could take a similar amount of time to analyze, but vary widely in another aspect. Also, generation and post-processing time can often be much smaller contributors to the overall duration than analysis time. There are always exceptions, which is why the durations should all be considered. When one set of data has a negligible contribution in comparison to another, the sets of data that do not provide helpful information for model selection can be ignored.

Checking for outliers can be done with varying degrees of sophistication, visually or otherwise, but the important part is that the duration used for model selection should be an accurate representation of the model effort instead of the idiosyncrasies of a particular machine. In cases where the model is inexpensive enough, repetitions can be performed, either a priori or after a cursory examination, to get a better time estimate.

5.7.3 Hypothesis

The main difference between the examination of required effort and the development of an understanding of model fidelity is that the durations do not need to be compared to experimental or validation data. They are only relative to themselves, so as long as similar equipment was used and they were examined for outliers, the values directly represent the computational efficiency. This leads to a more purely statistical analysis problem, and

can be aided by kernel density estimation in a different way than before. This leads to hypothesis 2.2.

Hypothesis 2.2 *If the generation, analysis, and post-processing costs are known, or the most significant subset of the three, then a score based on the cost per evaluation of all included models and multi-model gains or losses enables a more informed decision-making process.*

5.7.4 Model Cost Estimation

Given a set of sample durations for each model, the median duration can be found as an estimate of the cost. The median can be found directly from the data or as the median of a kernel density estimate given the durations as sample values. Unlike with typical model responses, the corresponding variable values that led to each sample can be ignored under certain circumstances. The location within the design space may have an impact on the required effort, which is why the variable ranges must be kept in mind, but this evaluation is looking at the model as a whole, so all of the samples can be weighted equally to estimate a distribution.

Outliers

As mentioned before, outliers must be considered since these values are experimentally generated on machines that may be performing tasks other than the analysis being timed. In this case, two simple outlier methods is used. The first assumes that the cost of a particular model should be somewhat normally distributed. The mean of the durations for a particular model are found, and values that are outside of 3 standard deviations from the mean are removed. The second outlier method depends on the interquartile range (IQR), or the distance from the 25th to the 75th quantile. The interquartile range is multiplied by a scaling factor, typically 1.5, and values below the 25th quantile $-1.5 \times IQR$ or above the 75th quantile $+1.5 \times IQR$ are removed. These two methods can be iterated until no more

outliers are found.

A simple assessment of the effectiveness of this method can be shown by comparing the median calculated directly from the data to the median calculated based on a kernel density estimate of the data. When no outlier method is applied, the difference in median approximations can be up to 15%. When even the simple outlier detection method described above is applied, this number drops down to ≈3%. This number could be brought down even further if more model durations were included.

Cost

A process similar to fidelity assessment could be put forth as before to generate the probability that a model is the most costly or most efficiency based on distributions of the sample times. However, these data sets are experimental values on a real-valued scale, specifically time, so it would be more useful to say that one model is 40 seconds more costly than another instead of 40% more likely to be the most costly. Using KDE to estimate the median of the costs makes the best use of the available data, but takes longer to evaluate than simply finding the median of a set of numbers. Density estimates, however, are useful for visual examination of cost data, especially for larger sets of data.

5.7.5 Cost Ratio Estimation

The estimate of a single cost number for each model provides insight into the model set similarly to the model probabilities in fidelity estimation. However, as before, further insight should be interpreted through analysis of combinations of models. For this, the work of Toal referenced in Section 5.5 is brought up once again[105]. It was mentioned previously that some of the recommendations put forth related to the ratio of cheap versus expensive evaluations used to develop a multifidelity regression.

The first metric used is the ratio between two models, C_r, where

$$C_r = \frac{C_c}{C_e}$$

with C_c meaning the cost of a cheap evaluation and C_e being the cost of an expensive evaluation. This assumes that the higher fidelity model is more expensive and the lower fidelity model is cheaper. While this is a typical assumption, it is not necessarily true in all cases.

While the cost ratios between any two models could be evaluated simply using the median costs, a more thorough process can be undertaken without adding an unecessary amount of effort by leveraging the process used in the comparative fidelity assessment for aligning model sets. The cost ratio between two models should be evaluated at specific design points, so if an estimate of cost is not available at a particular point in one of the two models, a linear interpolation is used to generate a time for comparison. Cost ratios should be with respect to the location in the variable space to account for cases when the model cost depends on the region.

Iterating through all of the pairwise combinations of models, the cost ratios at each design point are evaluated. This is done in both directions, C_i/C_j and C_j/C_i. Again, as with model cost, the sets of ratios are examined for outliers, and the estimated cost ratios between two models are taken as the medians of the remaining values. Also as before, the median of a kernel density estimate could be used instead of the direct median of the dataset when the data isn't very well behaved, but it will take longer, and there are cases where each is more effective at calculating two ratios that are the inverse of one another.

As mentioned above, this method of estimating cost ratio is used because it accounts for the possibility that design point location may impact cost. Since it is not simply a ratio of the cost estimates, there will be some uncertainty in the value found by multiplying a

cost by a cost ratio, i.e.,

$$C_1 \times C_r^{1,2} = C_1 \times C_2/C_1 \approx C_2$$

In models where the times have more noise, there will be more error in the cost ratio estimates. This is just something to keep in mind since reducing uncertainty in cost requires repeated runs of the same model, where evaluating any number of design points for a multifidelity model set is a difficult task. However, if a model is not automated to be able to evaluate different cases, but it capable of easily repeating the same case, uncertainty could be reduced for the cost and cost ratio estimates even if the model is deterministic. If the model is stochastic, repeated iterations are recommended anyway, so the costs of all runs should be recorded and used.

5.7.6 Multi-Model Combined Efficiency

The other metric used by Toal to represent rules of thumb for multi-model combinations is f_r, or the ratio of "expensive" evaluations that can be replaced by "cheap" evaluations[105]. It is reprsented as

$$f_r = 1 - \frac{n_{me}}{n_{ce}}$$

where n_{me} is the number of expensive evaluations in a multifidelity set, and n_{ce} is the number of expensive evaluations if the expensive model were the only model.

Importantly, the number of "cheap" evaluations is not the same as the number of "expensive" evaluations that are being replaced. Instead it is the number of evaluations that can be performed in the same amount of time. This can be found using C_r, e.g. if $C_r = 1/5$, 5

cheap evaluations could replace 1 expensive evaluation. The rules of thumb given are

$$f_r > 0.1 \tag{5.7}$$

$$f_r < 0.8 \tag{5.8}$$

$$f_r > \frac{1.75}{1 + \frac{1}{C_r}} \tag{5.9}$$

$$\tag{5.10}$$

Equation 5.9 is included to assure that more cheap evaluations are performed than expensive in a case where C_r is low enough that Equation 5.7 would recommend less cheap evaluations than expensive. Equation 5.8 enforces the logical assumption that replacing too many of the more expensive evaluations with cheap ones would degrade the accuracy of the surrogate, since there are not enough higher fidelity points to compare against.

Equation 5.9

$$\frac{1.75}{1 + \frac{1}{C_r}}$$

was semi-empirically selected over

$$\frac{1}{1 + \frac{1}{C_r}}$$

since using 1.0 was not a conservative enough recommendation. This makes sense, as $C_r = 1$ would result in

$$f_r > \frac{1}{1 + \frac{1}{1}} = \frac{1}{2},$$

saying that those two models make an efficient combination if at least half of the evaluations are replaced, even though the two models cost the same. With the other option, $C_r = 1$ states that

$$f_r > \frac{1.75}{1 + \frac{1}{1}} = \frac{7}{8},$$

or 87.5% of the evaluations would need to be replaced, which is above $f_r < 0.8$ in equation 5.8. This more clearly states that this combination does not boost the overall efficiency.

This can form the basis of estimating the efficiency of a multi-model ordered combination. To start with, the cost of evaluating cases from a set of models is additive. Therefore, the base estimate of cost for an ordered set of models is the sum of the estimated model cost.

$$\text{Startup cost: } \sum_i C_i$$

This is called the startup cost, as the user would intend to run at least one evaluation from each of the selected models. Because of this, the cost should be based not just according to the first model in the set, but the entire set. The efficiency of a model set is more complicated than that and based on the order in which the models are selected, similarly to the requirements for multifidelity model scoring. This is related to the relationship in Equation 5.9.

The definition of the cost ratio $C_r = C_{cheap}/C_{expensive}$ implies that this ratio is assumed to be a fraction in $(0, 1]$. This is further demonstrated in the efficiency ratio, since as shown above, when $C_r = 1$, the ratio shows that the model combination is too inefficient to be selected. Actually, any $C_r > 16/19 \approx 0.842$ yields a minimum f_r greater than 0.8. Conversely, as $C_r \to \infty$, the required f_r will only ever approach 1.75, which would not adequately penalize the efficiency of a poor multi-model selection.

As such, if a ratio similar to $\frac{1.75}{1+1/C_r}$ is to be used to show the efficiency of a good selection, when C_r is over a certain amount, the efficiency of the model set should be penalized by $\frac{1+C_r}{1.75}$. Since Toal implies that models with $C_r > 16/19$ incur some efficiency benefit, as described in the previous chapter, this should be the pivot point between the efficiency reward and penalty functions. When $C_r = 16/19$, $\frac{1+C_r}{1.75} = 1.05$, penalizing the combination by 5%.

Therefore, the piecewise efficiency equation should be represented as shown in Equa-

tion 5.11, and the function is shown on a log-log scale in Figure 5.25.

$$E_r = \begin{cases} \frac{1.75}{1+1/C_r} & C_r \in (0, 16/19) \\ \frac{1+C_r}{1.75} & C_r \in [16/19, \infty) \end{cases} \tag{5.11}$$

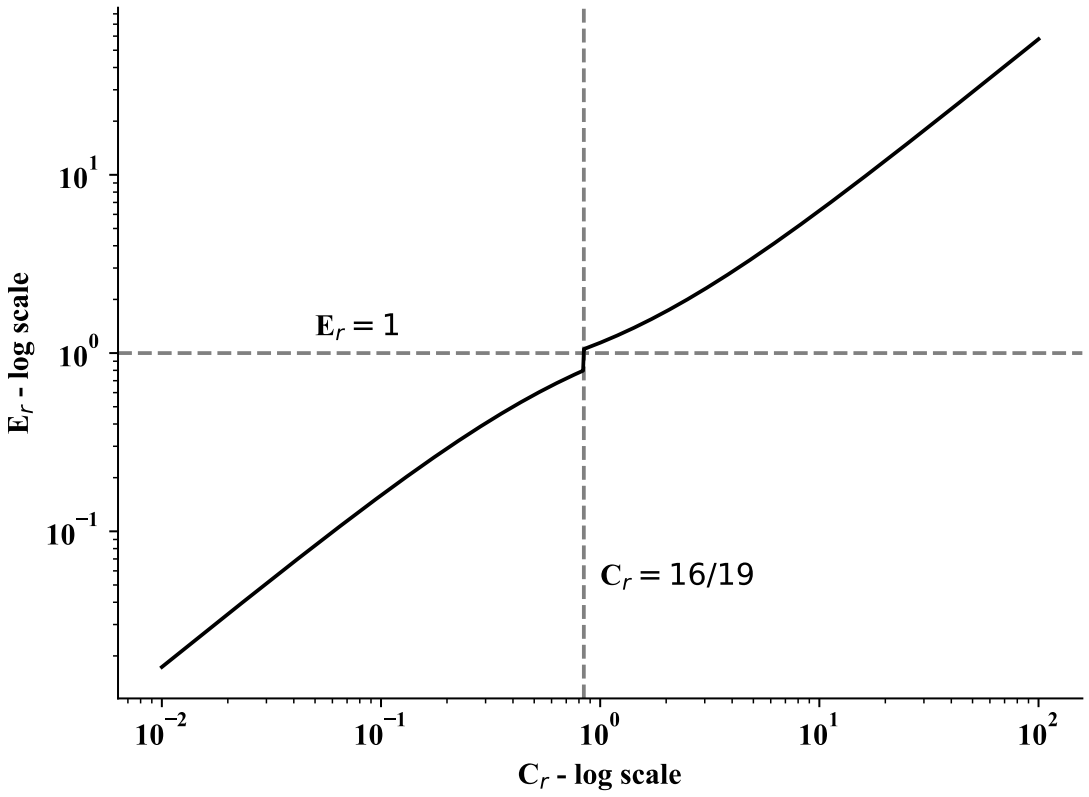


Figure 5.25: Piecewise Efficiency Scoring Funcation E_r

Based on this, an efficiency metric for an ordered model selection is the upfront cost, multiplied by the appropriate efficiency ratio E_r from 5.11, or, for M_1, M_2, ..., M_n:

$$\text{Set Efficiency: } \left[\sum_{i=1}^{n} C_i\right] \times \left[\prod_{i=1}^{n-1} E_r^{(i,i+1)}\right] \tag{5.12}$$

One thing that must be noted is that since this is measure of efficiency, it improves as it decreases, when the fidelity scoring metric improves as it increases. To evaluate a set of models in terms of efficiency alone will provide a ranking, but efficiency alone is never

the only metric. As such, the efficiency scoring metric will be evaluated further in the following section and compared with the assessment of fidelity.

5.8 Enabling Multi-Attribute Model Decision-Making

When a decision is based on one attribute, criterion, or objective, there is one theoretically optimal point that is better than any other. However, most engineering decisions are not based only on one dimension. In such cases, there is a set of points, called a Pareto front, that, instead of being optimal, are *non-dominated*. Such points are called non-dominated "in that no other set member exceeds a given design's performance in all goals[108, p. 179]." When the axes are model responses with respect to continuous design or uncertainty variables, there are many challenges to finding the Pareto front, as there are an infinite number of options. However, the number of single and multi-model ordered combinations is a discrete list, which may be large, as discussed in a previous section, but finite.

While a set of models can be evaluated in terms of fidelity and efficiency individually, the real contribution comes in the combined assessment of both attributes. The selection of a model will be dependent on an acceptable level of fidelity. It may seem counterintuitive, but going to the highest fidelity possible is not typically the best option. A higher fidelity seems preferable because it is difficult to define the minimal acceptable fidelity, so higher fidelity seems the more conservative approach to guarantee meeting the minimum requirement. However, higher fidelity than necessary leads to increased dimensionality which can actually increase uncertainty, as well as appreciably exacerbating the developmental requirements and evaluation cost.

In fact, minimum acceptable fidelity is a term highly coupled with accreditation, and, as such, is a subjective assessment. Defining appropriate resolution, abstraction, and scope depends on the significant contributors to a model's accuracy, which is where the experiential knowledge of experts comes into play. However, this experiential knowledge typically

is typically developed under particular circumstances at a certain point in the design process. This can sometimes cause a subject matter expert to completely discredit a model that does not account for a detail or phenomenon that has proven important to them in the past, regardless of the purpose of the current model being selected.

Therefore, this scoring system for fidelity and efficiency aims to help visualize the relative capabilities and limitations of the current model set, so that some of the subjectivity in the discussion of model selection can be removed. This is one of the general goals of data visualization, but providing a quantitative justification for discussions pertaining to fidelity is especially important in the early stages of a project.

5.8.1 Model Set 1: Notional Model Set

Single-Model, Multi-Attribute

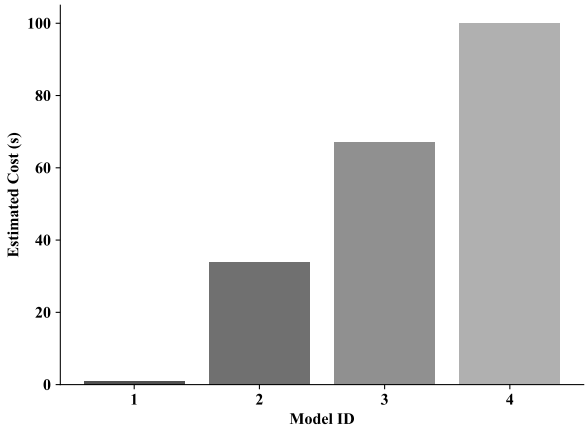
Looking back at the set of four notional models, the fidelity and efficiency scoring metrics can be tested for a variety of conditions. The three sets of notional costs are shown in Figures 5.26a, 5.26b, and 5.26c. The costs in Figure 5.26a follow an even linear spacing from 1 second to 100 seconds. In this case, some of the models are much cheaper than others, but the adjacent model is not an order of magnitude improvement in cost.

The second set in Figure 5.26b are evenly distributed on a log scale (1, 4.64, 21.5, and 100). This is a more ideal circumstance, such that the models other than the highest fidelity are much more efficient. Presumably, if the option most likely to be the highest fidelity is too costly, any multifidelity combination would significantly outperform the single model option, as will be shown in the following section.

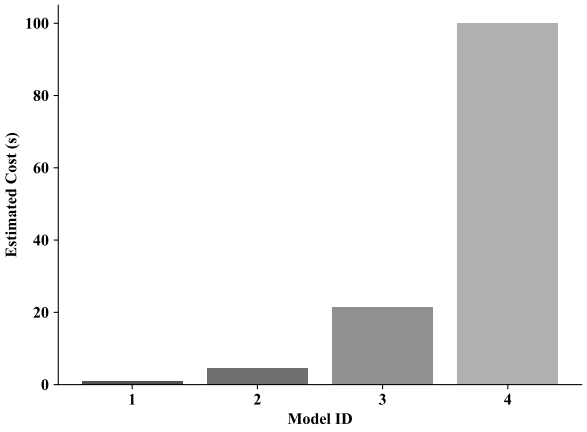
The third option in Figure 5.26c, while it may be difficult to tell, follows a linear progression as well, only from 100 to 101 seconds. In other words, the times are very similar. When the costs are very similar, it would logically be assumed that efficiency has much less to do with the decision-making process than fidelity, as multi-model combinations do not lead to a more efficient collection of evaluators. In such cases, the only reason for selecting

multiple models is for robustness, when applicable.

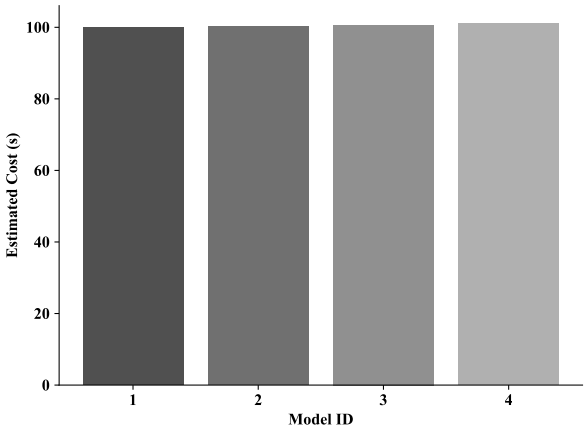
It is worth noting that in each of these cases, the cost is assumed to increase, even if very slightly, from the lowest fidelity model to the highest. This is, as mentioned before, often the case, but not generally true. However, it is valid to use here, as a lower fidelity, higher cost model would simply be dominated by another option.



(a) Linear Progression



(b) Geometric Progression



(c) Similar-Valued

Figure 5.26: Notional Model Cost Scenarios

Using the probabilities of highest fidelity shown in Figure 5.5a and a linear progression of model costs, the Pareto front for the individual models are shown in Figure 5.27. As described in Section 5.6, the score for the single model most likely to be the highest fidelity

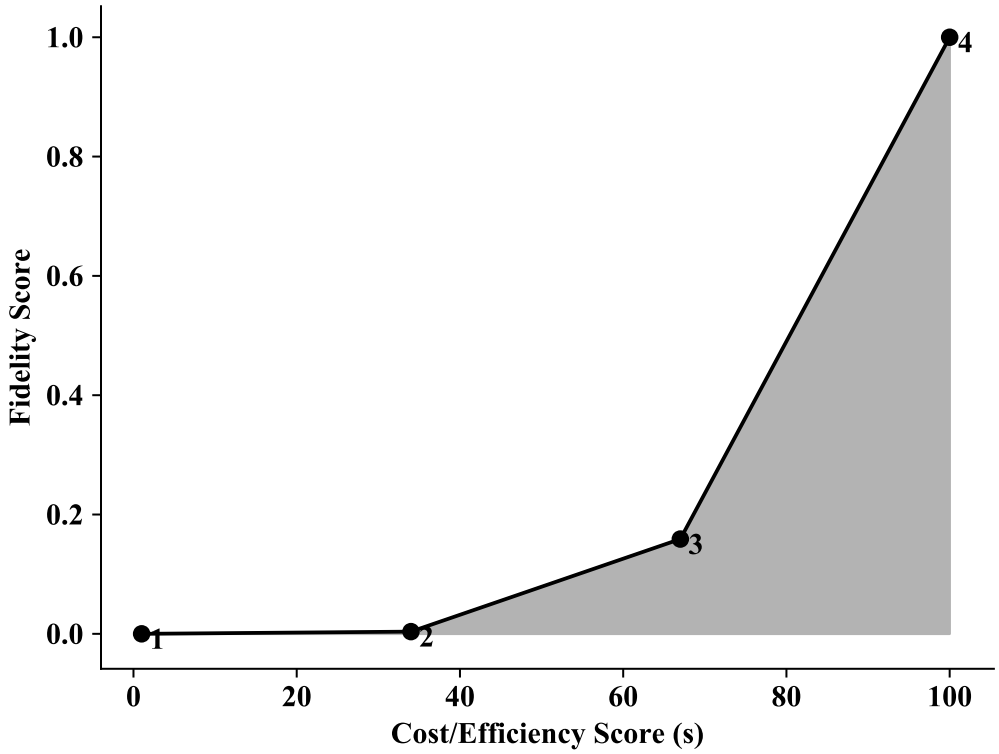


Figure 5.27: Notional Single Model Pareto Front: All Increasing Fidelity Attributes, Linear Cost Progression

is 1.0, and all fidelity scores are in $(0, 1]$. Also note that the fidelity score should be maximized, while the cost/efficiency score is minimized, so the shaded region to the lower right represents the region of dominated solutions.

If a cost of 100 per evaluation is too expensive for a certain scenario, cost requirements can be placed directly on this graph as a bound on the x-axis to show which option or options represent the current best selection. Given the fidelity scores when resolution, abstraction, and scope are all increasing, the fidelity score for models 1 and 2 are very low, which could lead to a fidelity bound on the y-axis representing a minimum fidelity requirement. As discussed before, however, this is a more subjective requirement, but could be justified in this case since the fidelity scores of those two models are nearly zero.

If, for example, model 4 is deemed too costly, then multifidelity collections can also be considered using the scoring methods developed in this chapter. There are many ways to find a non-dominated set of points. As discussed previously, the problem is simplified somewhat since there is a finite number of options. Specifically, when there are 4 models, there are only 60 multifidelity options.

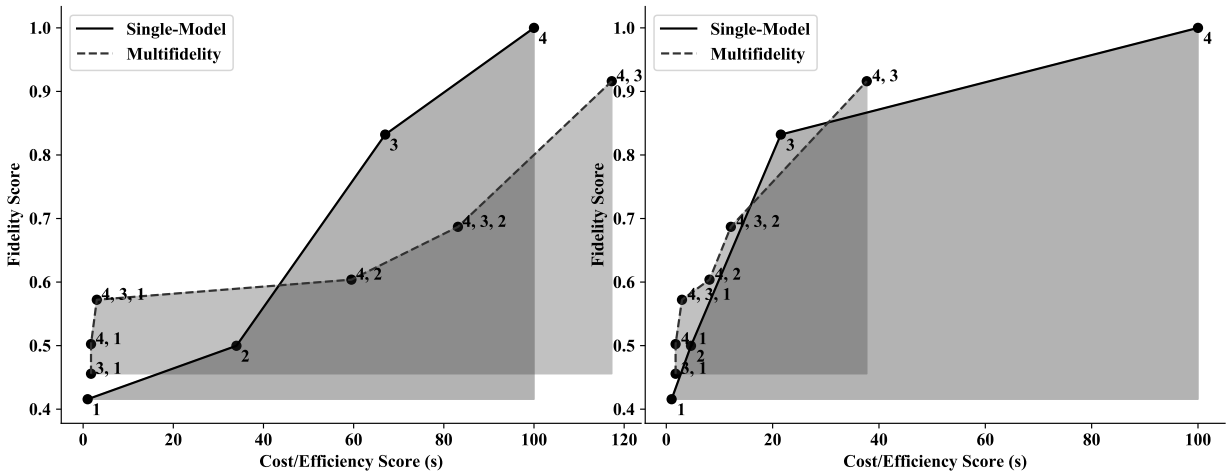
To find the non-dominated set from all of the possible options, the $non\_dominated\_front\_2d$ method from the $PyGMO$ Python package is used. $PyGMO$ and the associated $PAGMO$ package in C++ is an open source "scientific library for massively parallel optimization[109]." For more than two attributes, a different method must be selected, but this function is chosen when applicable as its complexity is $\mathcal{O}(N \log N)$, making it more efficient than more general options. For more information, see the work of Jensen[110].

If only the Pareto optimal, or non-dominated, permutations are saved, then the memory requirement issue in single-attribute scoring is alleviated. Of the 60 options, only 6 are non-dominated. As the model size increases, presumably, the number of Pareto optimal configurations will not increase as rapidly as the number of possible options. This will be proven out using the other model sets.

As discussed before, when there are 10 or more models, the number of permutations is above one million. If, for example, the non-dominated algorithm is run after every million permutations, then the memory requirement will no longer be constraining, and it will be a matter of evaluation time. Additionally, since, as discussed in earlier sections, larger permutations are generally less desirable, a stopping criterion could be put into place to limit the evaluation time as well. An example of this would be to stop evaluating new options once the number of non-dominated solutions falls below one for every million scored.

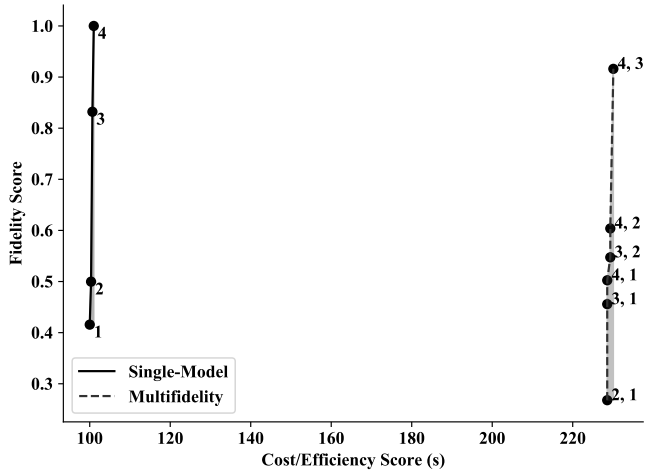
Even for a set of 100 models, the number of 1, 2, and 3 model permutations is 980,200, so they would all be examined. There would, however, be over 94 million 4-model permu-

tations. This means that the stopping criterion in the previous paragraph would likely come into play before reaching combinations of 5 models. This would, however, be preferable to waiting for all $2.537e + 158$ multifidelity permutations to be scored, as the development and implementation of anywhere near 100 models is always going to be excessive.



(a) Linear Costs

(b) Geometric Costs



(c) Similar Costs

Figure 5.28: Single and Multi-Model Pareto Fronts: Realistic Fidelity Attributes

Changing from the "All-Increasing" fidelity assessment to the "Realistic" option shown in Figure 5.5d, the lowest single-model fidelity score increases to 0.42, with the highest single-model fidelity score still at 1.0. The non-dominated multi-model options are shown

in Figures 5.28a, 5.28b, and 5.28c.

For the linear progression of cost in Figure 5.28a, as expected, the models that are the second and even third highest in terms of fidelity are too close from a cost standpoint to provide enough of an improvement in efficiency to dominate the single models. Only the permutations that include the cheapest model are favorable to a single model option. On the one hand, model 1 is the least likely model to be the highest fidelity, though on the other hand a fidelity score above 0.4 may be acceptable. As the fidelity and cost both increase from model 1 to model 4, the scores also favor the ordered collections that move from model 4 downwards, since this is the "correct" direction on both axes.

Alternatively, in Figure 5.28b, when the costs follow a geometric progression, meaning model 4 is over 4.5 times more costly than even the adjacent model, the multifidelity Pareto front lies entirely in the single-model non-dominated region. When model 4 by itself is too costly per evaluation at 100 seconds, using models 4 and 3 reduces the estimated cost below 40 without much of a drop in fidelity. Additionally, the combination of 4, 3, and 2 reduces the cost metric well below 20 while the fidelity score is only down to ≈ 0.7. Model 3 by itself is still worth consideration, but this is the optimal type of situation where multifidelity methods could be applied, assuming model 4 is too costly.

The exact opposite could be said about Figure 5.28c: the case where all of the models are similar in cost. The combination of any two models will approximately double the standup cost, without giving any efficiency reward in the process. Additionally, only two-model combinations show up as non-dominating, as the affect of standup cost greatly overshadows the other effects.

However, it should be kept in mind that cost is always relative. If double of even triple the standup cost is still well within the compuational budget, then multiple models could be evaluated for a more robust solution, keeping in mind that there is no benefit from an efficiency standpoint.

Conclusions

The evaluated cases reiterate that the fidelity and cost scoring methods meet the necessary requirements, as a single-model Pareto front can be generated and compared to multifidelity options. Multi-model combinations that are selected in decreasing fidelity and cost score more highly, and, given enough of an efficiency benefit, dominate the single-model combination. Multifidelity scores represent a combination of the fidelity score of the included models, taking into account whether the models were selected in the best order possible. When the costs are similar, the initial cost dominates, meaning that multiple models represent a poor selection from a cost standpoint.

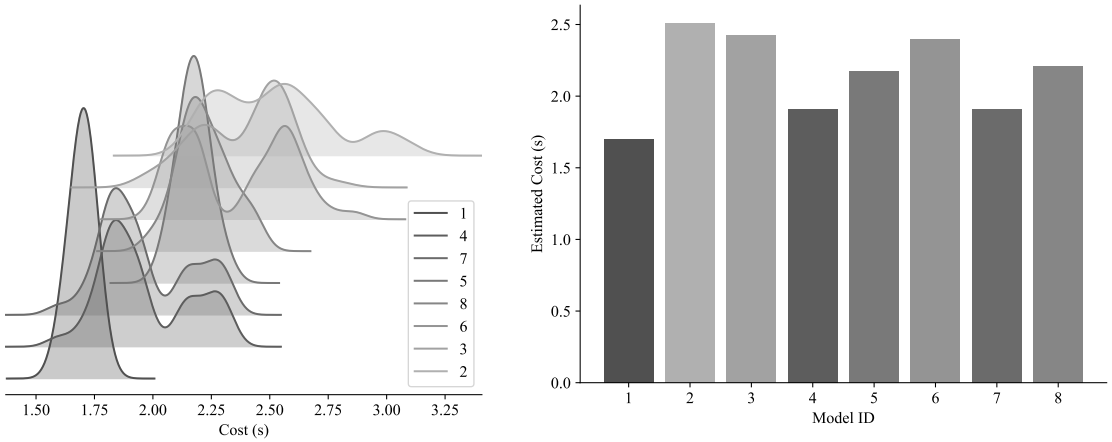
5.8.2 Model Set 2: I-Beam FEM

However, these models, their fidelities and associated costs are still notional. To further test out the scoring methods and Pareto front generation, model set 2, the I-beam finite element models, are evaluated and compared. Specifically using the adjusted probability of highest fidelity using the linear buckling results, as shown in Figure 5.19, the fidelity and cost scores are calculated.

The distribution and median costs for the down-selected set of 8 finite element model options is shown in Figure 5.29. Notice that while there is some variation, all of the costs are between 1.5 and 2.5 seconds, meaning that not only are they efficient, they are also similar in cost.

The Pareto fronts for single models and multifidelity combinations are shown in Figure 5.30. As discussed above, when a model does not fall in the same order for fidelity and efficiency, it falls into the dominated region of the single model Pareto front, as evidenced by models 2, 3, 4, 5, and 6. Models 1, 7, and 8, however, form the fidelity/efficiency single-model Pareto front.

The multifidelity options in this case are also more interesting than the notional case. Model 6 has a high enough fidelity and cost such that the multifidelity scores are favorable.



(a) Distributions with outliers removed

(b) Estimated costs

Figure 5.29: Costs For Down-Selected Set of 8 I-Beam Finite Element Models

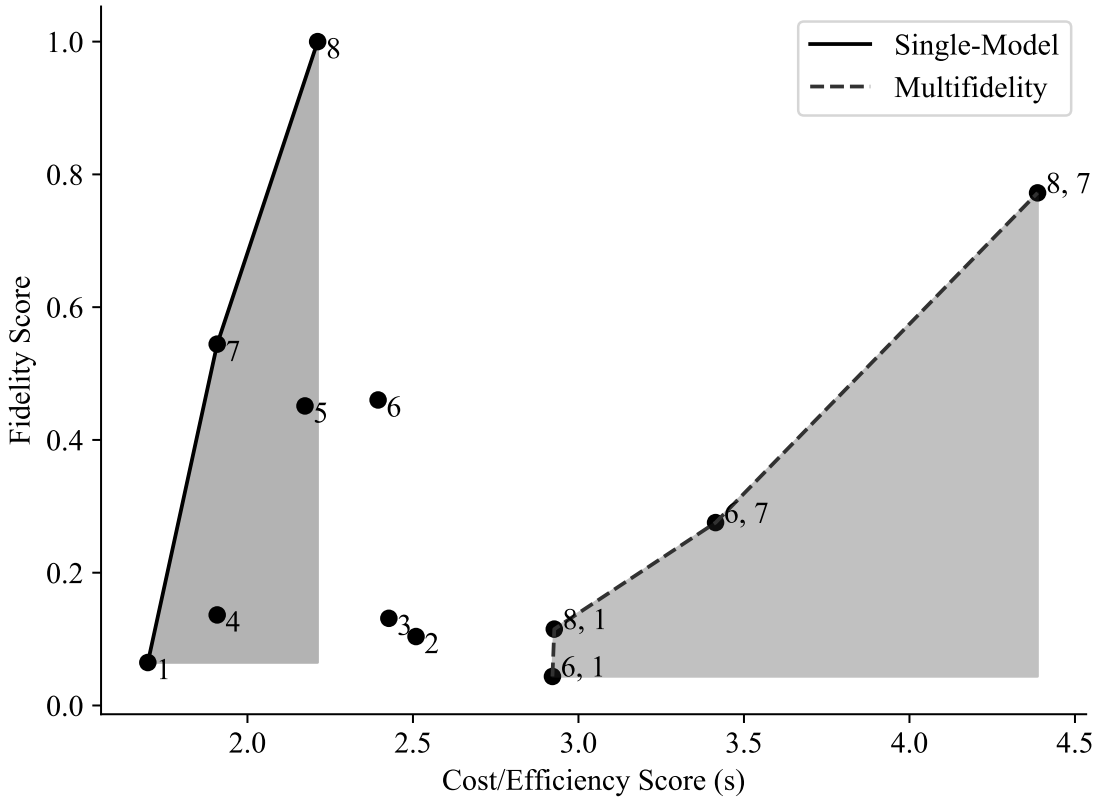


Figure 5.30: Single Model and Multifidelity Pareto Fronts for 8 I-Beam FEM Set

Specifically of note, the set of models 6 and 7 are on the multi-model Pareto front. This is due to the fact that the fidelity for 6 and 7 is similar, with 7 being higher, but since model

6 is noticeably more expensive than 7, the non-dominated combination is the one moving from model 6 to 7. The efficiency boost outweighs the difference in fidelity ordering.

Additionally, model 6 comes up again the combination of 6 and 1, since 6 is significantly higher in terms of fidelity, and sufficiently higher in terms of cost to present itself as a good combination. The other non-dominated multifidelity collections, 8-7 and 8-1, follow the expected pattern of combining models from the single-model Pareto front. It is worth noting that due to the fact that all of these models take less than 5 seconds to run, the difference in estimated cost between them could largely be due to noise, so the expected efficiency benefit from combining two models may not be fully realized under further repeated evaluation.

Since the models are similar in cost, the multifidelity Pareto front falls within the overall dominated region. However, since the models are efficient, the estimated csts are still small enough that multiple models could be carried forward if the cost per evaluation requirement allows. By continuing to develop multiple models, the user would be hedging their bets against currently unrepresented behavior. In terms of fidelity, the score for model 1 might be lower than deemed an unecessary risk, and left out of future development.

5.9 Conclusions

In this chapter, the primary methods of this thesis are developed. The fidelity framework developed in Chapter 4 is used to clarify the questions asked of experts with respect to model fidelity. The fundamental characteristics of resolution, abstraction, and scope are used for qualitative assessment as they are easier to understand and less subjective as opposed to the single metric of fidelity. Descriptive fidelity assessments are to be performed in terms of orders instead of explicitly-defined values, as the exact ratios between models are difficult for experts to define in a justifiable way.

From there, the scores generated using the qualitative orders provided by experts with respect to the three aspects are combined using KDE to generate a probability distribution

estimate for the relative level of confidence in the fidelity of each model in the set. Comparison of these distributions can be used to calculate the probability that a given model is the highest fidelity available. The probability of coming in second through last can also be found, which helps to provide fidelity insight.

To this point in the methods, the understanding of the model set is entirely data-independent, meaning that insights can be visualized regarding modeling options without having to wait for lengthy development cycles. If, however, model data is available, methods are developed to make use of that data through comparative data analysis. The correlations and errors between the comparable data sets are used to show how much the models agree with each other. Given that models of different resolution, abstraction, and scope are estimating the same response, agreement between these approaches is treated as an indicator of higher fidelity. While it was asserted that experts should not be required to define the specific ratios between models, model data can be used to adjust the relative magnitudes in a meaningful and traceable manner.

The relative cost of different models requires evaluation of the models, but not to the same extent as for response prediction. Each evaluation provides a new experimental value that can be used to estimate cost. While the cost could vary somewhat depending on the location of the design point, for a reliable model within valid ranges, variation should not be significant and cannot be generally assumed to occur. Therefore an estimate of cost can be made based on even a single value for model generation, analysis, or post-processing time.

Based on the probability of a model being the highest fidelity in the set, and estimate of cost, and the estimate of relative cost ratios between models, fidelity and efficiency scores can be generated using the methods developed herein. From this, the non-dominated set of single and multiple models can be found. Minimum fidelity requirements, however subjective, could be applied to the Pareto front if they exist, as well as requirements for cost per evaluation, to determine the most appropriate model or models for the current

point in the design process.

A notional set of four models was useful for testing out how fidelity probabilities and scores are generated with respect to a descriptive assesment of fidelity. The notional set also allowed for exploration of how the single a multiple model Pareto fronts compare for different relative model costs. However, the comparative data analysis techniques could not be tested using this set as there are no actual models to generate predictions, and the generated costs are purely notional.

To generate a set of model data based on realistic models, a set of I-beam finite elements were developed and tested for three different problem definitions to provide varying set of data. Model set 2 allowed for a trial of defining the order of models in terms of resolution, abstraction, and scope, and how the descriptive probability of highest fidelity could be generated. Following that, the comparative data analysis methods were tested using the linear static, linear buckling, and normal modes responses.

The linear static responses showed high levels of agreement, showing how models that provide duplicate responses can be filtered out of the set to simplify model selection. The linear buckling responses showed some agreement, but also distinctly different trends. The correlation and error metrics used to assessment model agreement also showed the capability of the methods to point out models that either need troubleshooting or do not adequately capture the appropriate phenomenon for the defined problem. Specifically, models that performed well for the linear static problem needed changes to the boundary conditions to be usable for linear buckling. Since these were left out, the responses did not agree with the other models, showing not only that the developed method was capable of finding these issues, but illustrating the danger of saying "we're going to use this model becuase we already have it" without thoroughly investigating applicability to a new application. The normal modes results mostly agree with one another, except for one model. However, unlike with the linear buckling response, the difference appears to be do to the representation of the phenomenology of the problem.

The fidelity and efficiency scoring methods are also tested for the I-beam FEM set, showing that since they are all very efficient, the model decision-making process depends more on fidelity than cost. In addition to the uniformly low cost, model set 2 represents a simple, fixed scope set of highly reliable models. As such, it would be beneficial to further test these methods for more complex models, with varying costs and scopes. While it has been shown that leaving out scope, even when it is the same for all models, represents a potentially significant difference in fidelity assessment, a multi-scope model set is preferable for providing further justification for the inclusion of scope in the fidelity framework. Therefore, the next chapter uses the methods developed in this chapter, applied as a framework for enabling informed model decision-making, to the more realistic trade study of how the estimated structural wing weight of an aircraft varies as the wing aspect ratio deviates from the baseline.

CHAPTER 6
AIRCRAFT WING WEIGHT USE CASE

6.1 Introduction

In order to show how the fidelity framework developed in Chapter 4 and the methods developed in Chapter 5 can be applied to a realistic model selection problem, an aircraft-related trade study is selected. The steps that must be undertaken to enable informed decision-making in terms of fidelity and efficiency are described in this chapter for this use case.

The steps that must be followed are similar to that of any generic decision-making process. Specifically, they are similar to that of the generic Integrated Product/Process Development methodology developed at Georgia Tech [111]:

1. Establish the Need
2. Define the Problem
3. Establish Value
4. Generate Feasible Alternatives
5. Evaluate Alternative
6. Make Decision

For the problem of decision-making from a set of multifidelity options, the steps are redefined as:

1. Problem Set Definition
2. Model Set Development
3. Descriptive Fidelity Assessment

4. Adjusted Assessment Given Model Data

    (a) Using Model Data to Identify Deficiencies

5. Fidelity and Cost Scoring for Multi-Attribute Decision-Making

6. Iterating as Data is Generated and Requirements Change

This framework is to be enumerated and used to test out the methods of the work in this chapter, starting with the first step, problem set definition.

6.2 Step 1: Problem Set Definition

Problem set definition, akin to establishing the need and defining the problem, is an important step that must occur before the methods of this work can be applied. The problem is one that needs a model to solve it: designing a system, evaluating improvements, performing a trade study, or other specific analysis or optimization tasks. For the linear static deflection of an I-beam, this could have been finding the appropriate length of a beam that would deflect less than available clearance. The linear buckling problem could involve a need to determine the minimum acceptable length that would withstand a pre-determined load. Normal modes analysis can be used to help find a structure that will not be affected by the ambient frequencies.

The aircraft use case to be explored in this chapter entails wing primary structural weight estimation for deviation from a baseline aspect ratio of an aircraft outer mold line (OML), which will be defined in more detail below. The estimation of aircraft wing mass is a common aerospace problem. Some of the complications of this process were described earlier in Section 4.6 while defining the fidelity framework. Aircraft structural design is to be used here to represent a more realistic problem for development of a multifidelity model set. The models included herein are not intended to represent a comprehensive list of all possible manners for estimating wing weight. Instead, it is simply a subset of varying

fidelities developed based on the authors experience and an understanding of some of the common preliminary methods used in the industry and in published literature[4, 5].

This defines the point in the design process for which the decision-making process will be occurring. The Manufacturing Influenced Design methods referenced attempt to bring models forward to allow for understanding of process-based manufacturing inputs and parameters in preliminary design. This means that many of the initial design decisions have been made about a vehicle: configuration, payload type and quantity, etc. However, since aspect ratio is still a fairly high-level parameter to be varying, this presents challenges in terms of resolution, abstraction, scope, and efficiency.

6.2.1 Vehicle: NASA Common Resarch Model

The design space selected for this use case is a modification of an existing vehicle definition. Using an existing vehicle description provides information such as an outer mold line, meaning many of the design decisions have already been defined in the conceptual design phase. Specifically, the aircraft used as the baseline here is referred to as the NASA Common Research Model, or CRM[112]. The common research model was developed by NASA as a platform for publishable research. The geometry is representative of a modern transport aircraft similar in scale to a Boeing 777. A general description of the aircraft can be found in Table 6.1 and a three-view is shown in Figure 6.1.

Most of the initial work regarded gathering aerodynamic data through scale wind-tunnel tests and computational methods and making the results publicly available. However, quite a bit of work has been done and published in disciplines other than aerodynamics, as this platform makes it easier to find a wide array of non-proprietary information about a realistic baseline aircraft. As an example of the aerostructural work that has been done is that of Kenway, Martins, and Kennedy[113]. They added representative internal structure and used optimization techniques to derive the undeformed shape of the lifting structures since the geometry used for wind tunnel testing is based on the deformed, in-flight shape.

Table 6.1: Common Research Model General Wind-Tunnel Model Description

Parameter Baseline	Value (Imperial)	Value (Metric)
Mach Number	0.85	-
C_L	0.5	-
Reynold's Number (Re)	$40e6$	-
Aspect Ratio (AR)	≈ 9.0	-
Taper	0.275	-
Side-of-body	10% of span	-
Yehudi break	37% of span	-
Washout	8°	-
Leading Edge (LE) Sweep	35°	-
Planform area (S)	3.01 feet^2	2796.38 cm^2
Span (b)	62.46 inch	158.648 cm

An extensive list of publications based on the CRM can be found on the NASA CRM website[112].

Outer Mold Line Definition

The OML of the vehicle is defined and stored in an OpenVSP model. OpenVSP, which stands for Vehicle Sketch Pad, sometimes just called VSP, is an open source software package for the parametric definition of aircraft[114]. As the name implies, vehicle sketchpad allows for a vehicle concept to be quickly drafted up, and then the geometry can be exported for use in analysis. The API of OpenVSP is used to adjust the aspect ratio according to a given design point which is then exported to define the external geometry of the various models.

6.2.2 Trade Study Variable: Wing Aspect Ratio

The parameter being modified for this case study is the aspect ratio of the wing. Aspect ratio is a term used in a number of contexts, but in an aircraft wing, it refers to the ratio of the span to some measure of the wing area. If the wing is rectangular, then it is simply the ratio of span and chord, otherwise it is the ratio of span squared to the projected area.

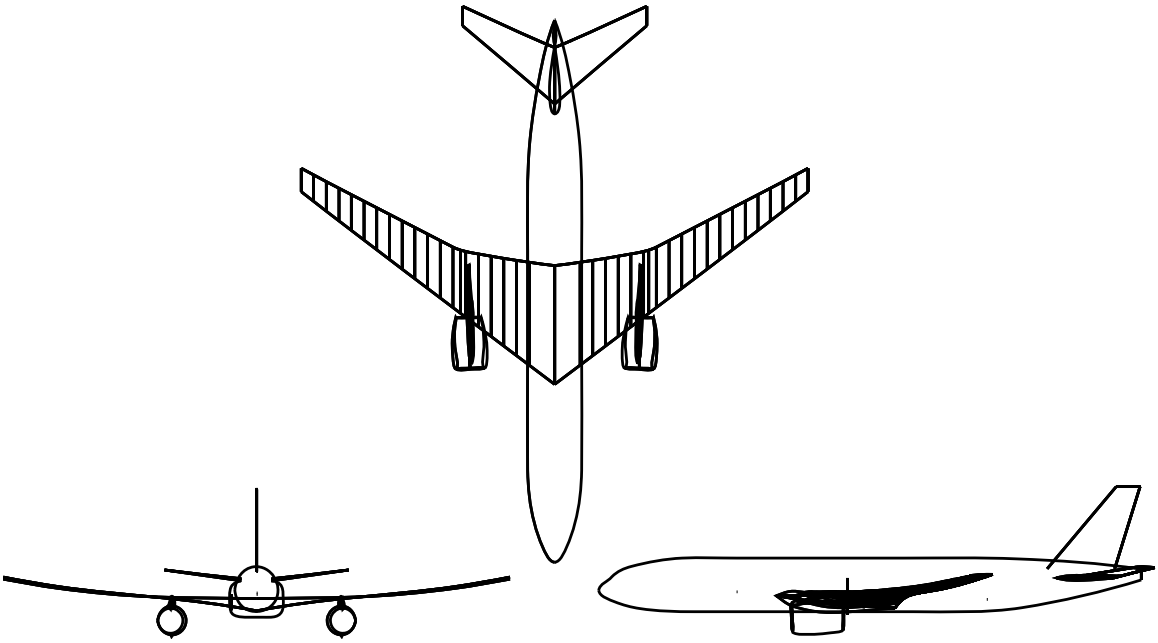


Figure 6.1: Top, Front, and Side View of CRM Geometry

More information about aspect ratio can be found in Raymer's aircraft design text, among others[106].

Wing aspect ratio was selected as the variable of interest because it impacts a number of aircraft design variables. High aspect ratio wings are more aerodynamically efficient, but there are a number of drawbacks. One of the main drawbacks for a commercial aircraft is that above a certain span, the aircraft will not fit in a standard airport terminal. In those cases, some sort of folding wing design must be used to allow for ground equipment clearance, and the hinges of such a design tend to drive up the wing weight. However, these types of considerations will be ignored here for simplicity.

A trade in the aspect ratio is of interest here since while aerodynamically, a higher aspect ratio is more efficiency, but structurally, the opposite is true. A short, stout, wing would be much easier to design in a way that will be light yet relatively rigid. As the aspect ratio increases, knowing that mass reduction is critical, the long, thin-walled, stiffened structure will be much more flexible from root to tip. Excessive flexibility in the structure can make the vehicle difficult to operate near the ground, or simply be a detriment to the

aerodynamic capabilities of the designed outer mold line.

An example of where this comes up in the comparison between two extremes of aircraft configuration: gliders and fighter aircraft. Gliders are designed to be as aerodynamically efficienct as possible, since they do not have engines to help move air over the lifting surfaces. Additionally, because they are not powered, they are going to be moving at comparably low speeds. High aspect ratio wings, from both aerodynamic and structural standpoints, limit the maximum recommended velocity of the aircraft. Interference with shocks and a dramatic increase in drag could cause very high aspect ratio wings to be very difficult to design for transonic and supersonic travel. This is one of the main reasons fighter aircraft have low aspect ratio wings. They perform better at higher speeds, and are more rigid to resist the high speeds and intense maneuvering required of a fighter aircraft.

The detriment due to the flexibility of high aspect ratio wings comes primarily from what is called the aeroelastic effect. As the name implies, it is the interaction between aerodynamics and the elasticity of the structure. If the wing is long and thin, there will be a great deal of deflection under load. As the wing deflects, it diverges from the ideal aerodynamic shape. This becomes a problem in analysis phases because as the structure is designed to a set of loads, the deflection of the structure changes the loads, and the process must iterate. Correspondingly, for a low aspect ratio wing that with less displacement, the aeroelastic updating will be less important, and vice versa as the aspect ratio increases. Therefore, the aspect ratio is an important consideration in model selection. More information regarding aeroelasticity in finite element modeling can be found in *Finite Element Multidisciplinary Analysis* by Gupta and Meek[79].

Importantly, the aspect ratio does not represent a simple, single change in the internal structural layout. While the span and chords are being modified, it is important to keep the taper, sweep, washout, airfoil shapes, etc. constant, which is why they are defined relative to percentages of span or chord. The two spars will remain at 20% and 63% of the chord, which are typical locations to provide stability while leaving room for the forward and aft

control surface mechanisms.

The ribs, however, are held at a fixed spacing of 30 inches (≈ 76.2 cm) apart and yawed outward 30°. This requires an algorithm to place new ribs as the span increases. One of the main considerations is that a new rib should not be added if the new rib would be very close to the tip rib. Another rib should only be added once there is enough space that it will be beneficial instead of just adding weight. Wing ribs are used to maintain the aerodynamic shape of the wing surfaces, but also to reduce the skin panel buckling length. Placing a rib reduces the un-supported span of the thin skin panel, making it easier for the skin stiffeners to resist buckling. Since new ribs must be added discretely, it leads to an interesting side-effect; while the aspect ratio is a continuous variable, the discrete addition of ribs can add a stepwise characteristic to certain responses. The significance of that effect is dependent on the relative proportion of the weight accounted for by the ribs.

Design of Experiments

Similarly to the I-beam model set, since there is one variable, a design of experiments does not require much sophistication to cover to the design space. However, for generality, standard methods are used to generate a DoE. A single-variable design of experiments was generated once again using a three-level full factorial for the upper, lower, and middle values. Then a 50-point Latin Hypercube design was again used, leading a 53-point design to be applied to each model selection. The baseline (aspect ratio=9), minimum (aspect ratio=7), and maximum (aspect ratio=15) aspect ratio wings are shown in Figure 6.2.

Structural Characteristics

Materials constitute a difficult problem in model definition and selection. The individual properties form a continuous scale, but a different material is technically a discrete choice. The material properties are defined based on some understanding of the typical distribution of each property for a given material composition from a given manufacturer. The

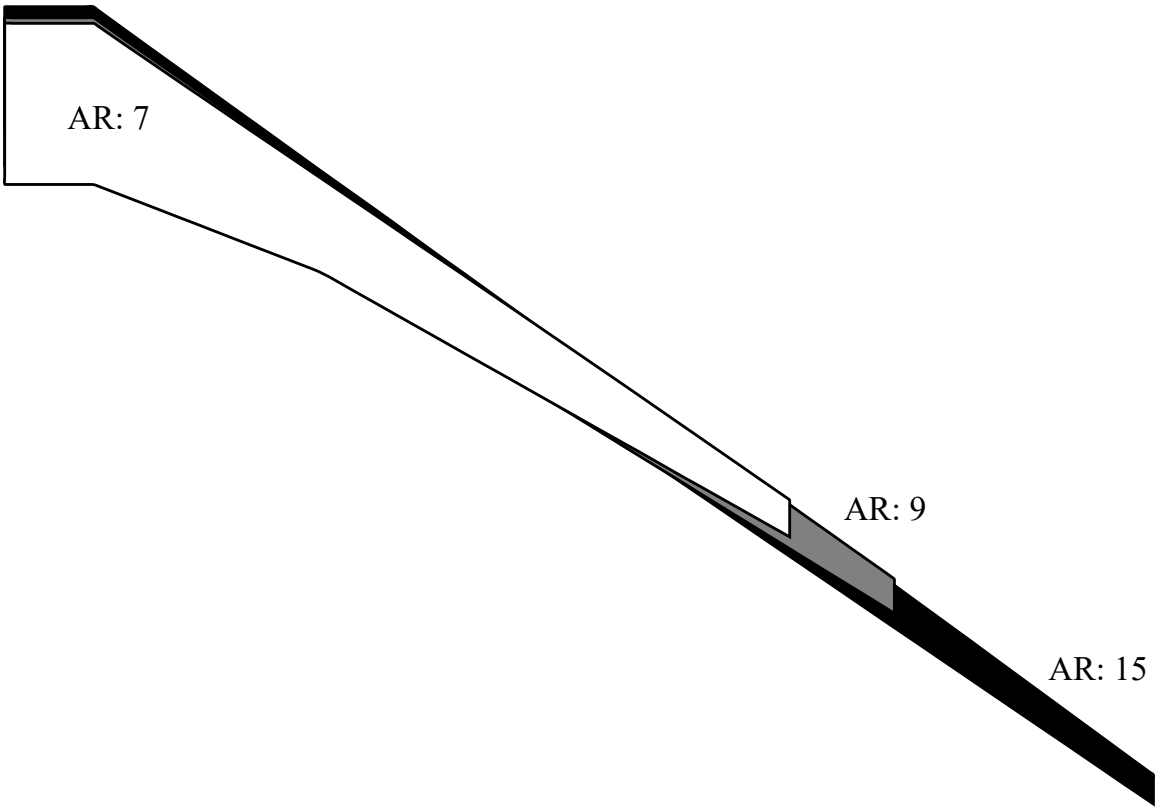


Figure 6.2: Baseline, Minimum, and Maximum CRM Aspect Ratio

structures of the models used in this set are made of an aerospace grade aluminum material properties, similar to 6061-T6, as described in Table 6.2.

Table 6.2: Aluminum Properties[115]

Property	Value (Imperial)	Metric
Density (ρ)	$0.0975\ lb/in^3$	$2.69879\ g/cm^3$
Modulus of Elasticity (E)	10.3e+6 psi	$71.016\ GPa$
Poisson's ratio (ν)	0.33	-
Stress limit in tension ($S_{tension}$)	47e+3 psi	$324.054\ MPa$
$S_{compression}$	43e+3 psi	$296.475\ MPa$
S_{shear}	30e+3 psi	$206.843\ MPa$

Additionally, and very important in structural modeling, the discrete change between a 6000 series aluminum and a 7000 series aluminum is not the same as the difference between a metallic and a composite. Structural optimization is based on constraints defined by the failure modes of the selected materials. Modeling becomes much more difficult

when moving from metals to composites. Metals are generally isotropic and homogeneous, meaning that there is uniformity and a lack of directionality to the material properties. The thickness of a piece of metal is essentially a continuous variable that only affects the properties in edge cases that can be handled with bounds.

However, composite laminates are made of discrete layers of material, with a certain thickness, that have directional material properties. This means that the thickness of the panel is now a discrete variable, based on ply thickness, and that at any point in a panel, the propreties are dependent on the number of plies, and the direction and order with which they are stacked. Additionally, thickness transitions require discrete "ply drops," holes for fasteners disrupt the properties differently than with metals, among other issues. In addition to the difficult of describing a composite structure, the failure modes or design constraints of isotropic materials are mostly based on stress, where those of composites are primarily based on strain.

These complications are mentioned to emphasize the point that material selection must be treated with care in model selection. Materials with fundamentally different design processes and failure modes, such as the difference between isotropic metals and anisotropic composites, essentially require separate model selection processes, which increases the work that is required.

Other structural parameters remain fixed for all models, some of which are described in Table 6.3. The engine location remains fixed under the wing. The fuel system is designed in three sections. The first section is in the center wing. The second section goes from side-of-body to 30% of the span. The third section goes from 30% of the span to 80% of the span. The fuel is assumed to remain full for all flight conditions. This is not the most realistic condition, but is conservative and simplifies the analysis.

Storing the fuel in the wing is convenient, since it is otherwise relatively unused volume, but it provides additional benefits. The mass of the fuel and the wing-mounted engine provide additional inertial relief for the wing. In other words, as the aerodynamic forces

lift the wing, the masses are pulling the mass of the wing itself, the contained fuel, and the attached engine in the opposite direction. This acts to counter the deflection caused by the lifting force, which means that the wing does not have to be as stiff. Leaving the fuel or engine out of the scope of the model would, consequently, change the results in an unrealistic way. As such, they are included for all models used herein.

Table 6.3: Fixed Vehicle Characteristics

Property	Value (Imperial)	Metric
Cruise Altitude	$37,000$ feet	$11,277.6$ meters
Cruise Mach Number	0.85	-
Maneuver Altitude	0.0 feet	0.0 meters
Takeoff Gross Weight	$650,000\ lb$	$\approx 300,000\ kg$
Engine Mass	$13,000\ lb$	$\approx 5,900\ kg$
Sizing Flight Conditions	$+2.5G$, $-1.0G$	-

6.2.3 Trade Study Response: Primary Wing Structural Weight

Primary Response: Sized Primary Wing Weight

As mentioned above, the primary response of interest is the estimated wing weight. Specifically, this is the predicted mass of the primary load bearing structures, defined as follows:

- Two spars
- Ribs
- Upper and lower skins between the spars

The fuselage of the vehicle is not represented in these models, which represents a scoping issue. The wing is connected to the center wingbox, which is where the boundary conditions are applied. In reality, this structure passes through and is attached to the fuselage, which would modify its stiffness. However, as the pass-through structure is intended to carry most of the load anyway, it is more conservative to ignore the structure of the fuselage.

Secondary Response: Cruise Wingtip Deflection

An additional response is tracked for each case for verification purposes. The cruise wingtip deflection is saved for each design point as it is a representative characteristic of the flexibility of the wing. Cruise tip deflection can be used to provide a quick verification that the material properties and loads are defined correctly such that the deflection isn't excessively large or small.

This can be a problem, especially when using an English unit system. A commonly used English unit system used in finite element modeling represents force as pounds-force, but also mass as pounds-force, such that density is pounds-force per cubic inch. In such a case, a Nastran parameter called *WTMASS* must be defined as 0.00259, or one over the gravitational acceleration, 386.09 inches per second squared[116]. If this parameter is missed, then the inertial forces will be off by a factor of 386.09, and the resulting cruise wingtip deflection returned by the finite element solver may be noticeably larger than the span of the wing, which is very unrealistic.

6.3 Step 2: Model Set Development

When a problem is presented that could be assisted by model data, a certain amount of expertise is needed to understand which models will suffice. This part of the process must always be somewhat manual, to determine which set of mathematical representations will not only provide the desired response, but account for the important characteristics of the problem. Over time, with data, it becomes more obvious just how sufficiently a problem aids in a particular problem, but initially, it requires a combination of expertise and research to determine which models have been developed or can be developed. Additionally, there are case specific issues, such as software licensing, that apply requirements to the model decision-making process that cannot be generalized. This framework seeks to understand those that can be generalized: the description of fidelity, analysis of available model data,

and understanding of model costs.

As models are being put forth for consideration, they may be at different phases in the development process. The best case scenario is models that are already developed and have been previously applied to the same problem, meaning they have been accredited, and should be verified and validated. Only the other end of the scale, a model could just be a concept, based on knowledge of modeling tools and mathematical representations. This means that the model's validation is purely based on expert opinion, and a full development cycle is required for implementation. In this case, the descriptive assessment of fidelity could be used to justify further consideration, and, as data becomes available, can be checked through comparison. Ideally, models that are to be considered have been through at least some amount of verification. This means there is the possibility that a limited data set could be generated, even if some of the process is still manual. This level of model development could allow for comparative data assessment to justify further development, act as a surrogate for initial validation, or to point out flaws that still exist in the model.

Aircraft structural analysis and design is a very active area of research. Estimating the mass and stiffness of aircraft wings is essential to designing an aircraft since these structures generate most of the lift, meaning they must maintain a very well thought-out shape, and carry a great deal of load in the process. Additionally, wings carry most of the fuel, and often the engines. A great deal of forces must be balanced, while striving for the lowest possible weight.

Models to estimate the wing weight of a design vary from regressions of historical data, structural models where the wing is represented as a beam, simple shell models, all the way up to detailed CAD representations of every flange, cutout, and bolt hole to find the exact mass and aid with manufacturing. The level of model used here in a fidelity-forward preliminary finite element design model. Shell elements are used so that the skins, ribs, and spars are represented, without locking the design into too many specifics. The design

of stiffeners, cutout, fasteners, and other more detailed features may be represented by the shell stiffness, but are not explicitly defined in the mesh.

6.3.1 Enabler: Rapid Airframe Design Environment (RADE)

The toolset used in this work to generate structural models based on a given OML is called the Rapid Airframe Design Environment, or RADE[117]. RADE has been developed over the past several years at the Georgia Institute of Technology Aerospace Systems Design Laboratory for the purpose of enabling preliminary-level analysis of parametric airframes. One of the basic requirements is that it be built on open source tools instead of requiring a finite element preprocessor, such as Patran, to be purchased. Because of this, most of the code is written in the Python language and leverages relevant available Python packages when possible. The initial description of the vehicle is often provided from OpenVSP, but since it is exported in a common CAD file format, it is not required to be from OpenVSP. However, VSP improves the efficiency of the process since the geometry it exports already has associated aircraft-specific metadata, e.g. the software already knows that a panel is a wing upper skin as opposed to just a surface.

RADE provides the ability to process the OML geometry, add internal structure, and generate a shell mesh for the intended analysis or optimization. While the toolset itself is developed with the open source concept in mind, the finite element solver is still MSC Nastran. This was selected because it is the industry standard, and allows RADE to leverage the aforementioned Nastran utility package developed by the author for the conversion of FEM data to Nastran-specific entities.

The parametric nature of mesh generation, flexibility in solution settings, and automation provided by RADE offer the ability to generate a variety of models quickly, instead of relying on a manual process in a graphical finite element preprocessor that can take days to generate a single model.

6.3.2 Introduction to Model Development Options

In addition to the reasonings described earlier, the optimization of aircraft wing primary structures is selected as a practical application since multiple fidelities of models are often generated for this analysis dependending on which aspects are considered most important. While an accurate understanding of the mass of the vehicle is crucial for performance estimation, many projects have succeeded or failed based the estimation of manufacturing costs. Modern materials and manufacturing processes make cost estimation based on historical data very difficult, but a more detailed cost based on the individual processes requires a great deal more data, specifically, a higher resolution description of the sized structure.

The methods for adressing this problem are sometimes referred to as Manufacturing Influenced Design, or MInD, some examples of which can be found in [4], [5], and [118]. The methods for efficiently sizing a structure in the appropriate way for process-based manufacturing analysis is part of what led to the initial codebase of RADE, and were developed, in part, by the author. One of the primary enablers for generating this level of structural detail is the HyperSizer software package[119].

HyperSizer

HyperSizer is a software package, similarly to Nastran, initially developed at NASA and then privatized by the Collier Research Corporation. Among its features, HyperSizer works as a post-processor and optimizer in conjunction with an external finite element model. One of its main strengths comes in the estimation of stiffness properties for a thin-walled, stiffened, structures as in an aircraft wing.

The method HyperSizer uses to generate structural properties while remaining flexible is referred to as the smeared stiffness approach. Representing panel stiffeners explicitly as mesh entities locks the model into a certain configuration. The way HyperSizer uses to work around this restriction is by designing the stiffener geometry using data from the

finite element model along with analytical equations, then calculating a representative stiffness for the finite element panel based on the stiffener configuration. These properties are applied simply to a section of shell elements, giving it realistic properties without the need for mesh adjustments.

This process was proven out by Collier research and shown in the following figure referenced from the HyperSizer documentation. The four mesh configurations were used to represent a rectangular wingbox-like structure designed under a simple loading. Representing all of the stiffener panels as shell is understood as the highest fidelity, but least flexible, representation, and is treated as a baseline for the comparison of calculated deflections and static design margins. The details are shown in table 6.4.

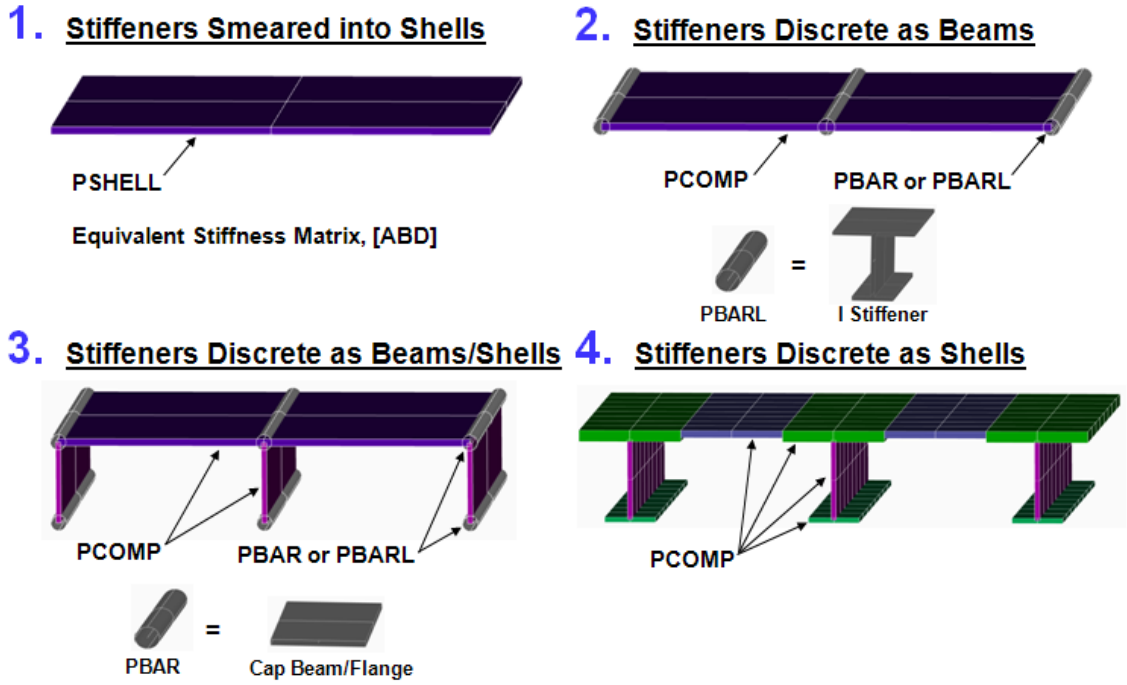


Figure 6.3: Examination of Smeared Stiffener Approach[120]

This points out an interesting example of the complex interplay between resolution and abstraction. While the beam and beam/shell representation reprsent a higher resolution in that the features are explicitly defined, the way that they are represented in the calculations is not only limiting from a meshing perspective, but less accurate. Importantly, the smeared

Table 6.4: Stiffener Representation Comparison[119]

Stiffener Representation	Deflection Error	Margin Error
Smeared	$+0.4268\%$	-0.08514%
Beam	$+2.985\%$	$+2.367\%$
Beam/Shell	$+2.407\%$	$+2.450\%$
All Shells	(Baseline)	(Baseline)

stiffener approach allows for access to all of the relevant attributes of the stiffened panel without requiring any mesh modifications.

6.3.3 Element Selection

Tri/Quad

The model set used here is defined based on a number of aspects not already mentioned. The first of which is the element type used to represent the surfaces. Here there are two options: purely triangular elements, or primarily quadrilateral elements with triangular elements only as needed in geometric transitions. Specifically, these are *CTRIA3* and *CQUAD4* Nastran elements.

It is easier to mesh the surfaces of the wing using triangular elements, but they are less accurate. Specifically, they are constant strain elements, and are known to be unrealistically stiff when used by themselves. Ideally, for accuracy, only quadrilateral elements would be used. However, quadrilateral elements work best when they are square. The curved, irregularly shaped, panels of a tapering wingbox would require some of the quarilateral elements to be skewed for full coverage. As such, it is acceptable to incorporate triangular elements as required to transition between sections. Both element types are included to show the comparison between ease of meshing and accuracy. For more information on the element types and their comparisons, see the Nastran Linear Static Users Guide[99].

6.3.4 Optimizers, Constraints, and Stiffener Description

While the element selection is independent of the other modeling options, the rest of the multifidelity characteristics are interdependent. The first set of options is based on the selection of an optimizer.

Nastran Optimization

The first optimizer category is what is built into MSC Nastran, selected by using solution 200. MSC Nastran uses either MSCADS or IPOPT depending on the size of the problem. MSCADS is the MSC-specific implementation of the open source Automated Design Synthesis code, written in Fortran. IPOPT stands for **I**nterior **P**oint **Opt**imizer, another open source code, that is more efficient for very large problems (more than $3000 \sim 4000$ design variables). IPOPT is more commonly used for topological optimization, so, while Nastran is allowed to automatically the select the optimizer, one of the algorithms in MSCADS should typically be what is used in this case.

When using the Nastran optimizer, one of two different optimization scenarios is selected. The first uses ten iterations of the fully stressed design algorithm (FSD) in Nastran. FSD is a relatively simple, efficient method for attempting to meet the design requirements in a traditional sizing routine. While the optimizer will be checking for convergence, it is not assumed to have converged in ten iterations. It is instead intended to act as a lower fidelity estimate, examining the physics and putting the design into a better position than the generic initial guess. This is a justifiable level of fidelity through comparison to works such as that of Courrier, et. al.[7]. In that work, the result of partially converged computational fluid dynamic (CFD) analyses were used as its own level in multifidelity regression. Despite not being a converged solution, allowing for a partial run of a simulation will provide an efficient guess at the behavior, and, as such, can be used to help interpolate between the points that were allowed to run to convergence.

The other scenario used with the Nastran optimizer begins with the same as before,

ten FSD iterations. However, it is followed up by traditional continuous, gradient based optimization iterations. The number of traditional design cycles is capped at 40 to prevent excessive run time if the optimizer is having difficulty. This is more than are typically needed to achieve convergence based on experience and recommendations in the Nastran documentation. For more information on the optimizers in Nastran, see the "Design Sensitivity and Optimization User's Guide[121]."

When using the Nastran optimization routines, the design objective is a minimization of mass. The internal variables are the thickness of each panels, and the design constraints are a simple stress constraint of each unstiffened panel. This means that the stiffeners are ignored entirely. However, the stiffeners are primarily there to keep the panels from buckling, and panel buckling is also ignored. Similar to the FSD approach, this is often a first step at approximating the weight, but cannot provide stiffener dimensions and is not assumed to provide the most accurate possible prediction.

HyperSizer Optimization

The third design scenario uses the optimizer in the HyperSizer software package. The optimizer in HyperSizer is a type of fully stressed design, and does not utilize gradient information. For each panel or stiffener variable, a certain number of discrete options is listed between the minimum and maximum values. The software then lists out all of the possible combinations in order by the resulting weight, and selects the minimum mass design that meets all of the constraints.

In this case, the sophistication of the optimizer is increased slightly through the following procedure:

1. Only allow three levels for each variable to reduce the number of total combinations
2. Perform the optimization
3. Based on which of the three values was selected, adjust the bounds for the following

iteration

4. Repeat the process until the maximum number of iterations or convergence

The bound-updating procedure adds an element of continuity back to the discrete selection of variable options.

While the optimization routine may not be as sophisticated as it could be, one of the main benefits of HyperSizer is that a wide assortment of failure criteria are built into the software. Deciding which constraints to use from the set of all available could allow for an additional level of modeling decisions, but for this case a common group of constraints for metallic materials are used for all of the HyperSizer models.

The other main benefit of using HyperSizer, as hinted at previously, is that the panel stiffeners are represented, through the smeared stiffening approach, both in the resolution and abstraction of the models. These models use an integral blade-stiffened panel concept. HyperSizer models can be used to provide an additional level of sized structural detail for detailed design or manufacturing analysis, as well as the higher resolution and lower abstraction estimate of weight.

6.3.5 Aerodynamics/Aeroelasticity

Another set of modeling decisions to be made relate to the aerodynamics and whether or not to iterate the loads and the structural optimization. The iteration of aerodynamics and loads is referred to as aeroelasticity and was discussed earlier in this chapter. In this case, ignoring aeroelasticity involves a straight-forward linear static structural analysis. Including aeroelasticity, specifically static aeroelasticity, updates the aerodynamic geometry based on the grid point deflections, re-calculates loads, and then runs performs structural optimization again.

Aero-static/AVL

The basic method of generating loads for these models uses the Athena Vortex Lattice (AVL) software[122]. The use of AVL for static load generation is one of the features developed as part of RADE. The aerodynamic surfaces are defined based on the geometry provided in OpenVSP. AVL trims the model and generates panel loads. Code built into RADE is used to convert those panel loads to forces and moments at a series of grid points running the length of the wing. Those grid points are near, but not coincident to, the ribs of the structure, and distribute the loads to the airframe using Nastran *RBE3* elements, which were discussed in Section 5.4.6.

Aeroelastic/Nastran

The evaluation of aeroelasticity utilizes another routine included in MSC Nastran, specifically solution 144. The aeroelastic routine includes a doublet-lattice lifting surface aerodynamic solver. The aerodynamic surfaces are defined in the Nastran input file as well as what parts of the structure mesh should be used for load transfer using splines. Specifically, the grid points at the intersection of ribs and upper skins are used to connect the aerodynamic mesh to the structural mesh. As the structural mesh deflects, the aerodynamic mesh is displaced accordingly. Then, the splines are used to transfer the loads generated by the aerodynamic solver back to the structural mesh. The benefit of using the Nastran aeroelastic solver is that once model is set up, the interim steps are handled internally.

However, the doublet-lattice method, unlike AVL, ignores the effects of camber, so some correction is needed to improve the accuracy of the aerodynamic solution. One of the methods for correcting this problem involves continuing to use AVL and the external method of transferring loads from AVL to the structural mesh sections via *RBE3* elements. The difference here is that all that is needed are the zero-degree angle of attack loads, since corrects, or shifts, the loads according to the shape of the airfoils through AVL. For more information on the doublet-lattice method or the Nastran static aeroelastic solution, see the

"Aeroelastic Analysis User's Guide[123]."

6.3.6 Scope

Wing, Wing-Panel Tail, Wing-Tail

The last modeling decision for the current set is dependent on the setting of the previous section. In the case that aeroelasticity is considered, the horizontal tail of the model must be represented in Nastran so that the solver can trim the aircraft. This is already a different scope than the rest of the models, since they only model the wing in Nastran. Typically, to reduce the difficulty of model generation and reduce the computational burden, the horizontal tail is only modeled structurally as a plate at the same location as the aerodynamic surface. The plate is defined to be made of a material that is incredibly stiff (modulus 100 times higher than aluminum) and very light (density of $1e-8$). This means that while the horizontal tail is included in the aeroelastic analysis, it is practically rigid and massless, and therefore the aeroelastic effects of the tail are essentially ignored.

To incorporate another more realistic level of scope, a second option is included which represents the horizontal tail more accurately. The same procedures are used for the horizontal tail as for the wing to build a flexible wingbox mesh according to the OpenVSP geometry. This means that the aeroelastic solver will have to account for deflections in the wing as well as the horizontal tail while iterating to convergence.

6.3.7 Model Set

Since most of the modeling decisions are independent of the element type selected, a shorter list can be written out of the possible combinations. The number of possible models in the set is duplicated between the all-triangular mesh option and the primarily quadrilateral mesh. The options are described in Table 6.5.

Note that there are nine different combinations, leading to a total of eighteen different modeling options once the difference between triangular and quadrilateral elements is in-

Table 6.5: Aircraft Model Set (independent of element type)

Optimization	Aerodynamics	Scope
Nastran-FSD	Static-AVL	Wing
Nastran-Traditional	Static-AVL	Wing
HyperSizer	Static-AVL	Wing
Nastran-FSD	Aeroelastic-Nastran	Wing-Panel Tail
Nastran-FSD	Aeroelastic-Nastran	Wing-Tail
Nastran-Traditional	Aeroelastic-Nastran	Wing-Panel Tail
Nastran-Traditional	Aeroelastic-Nastran	Wing-Tail
HyperSizer	Aeroelastic-Nastran	Wing-Panel Tail
HyperSizer	Aeroelastic-Nastran	Wing-Tail

cluded. However, as the models were generated and run, the Nastran-Traditional/Aeroelastic-Nastran/Wing-Panel Tail models seemed to take a very long time to run and did not produce good results. As these are still models that do not provide a detailed description of the stiffeners, they are not worth consideration unless they are more efficient than a comparable HyperSizer model. The results that were gathered will be discussed more later, but because of this, the similar models with a flexible tail (row eight) were not evaluated, leaving sixteen model types in this set, as described in table 6.6.

6.3.8 Importance of Scope

While the specifics of the responses of these models will be discussed in a later section, the multi-scope nature of this model set presents itself as an opportunity to further justify the inclusion of scope as one of the fundamental characteristics that drives model fidelity, along with resolution and abstraction. Multiple examples have been given to define how scope differs from the other aspects: how it is a boolean of whether or not part of the system is represented, since the propagation of resolution and abstraction effects cannot occur if the system attribute is not represented at all. Additionally, Observation 1.1.4 describes how much error can be incurred in the fidelity assessment process if an important aspect of fidelity is omitted.

Using the wing weight estimates provided by the various models in this set, a type of

Table 6.6: Aircraft Model Set

ID	Topology	Optimization	Aerodynamics	Scope
1	Tri	Nastran-FSD	Static-AVL	Wing
2	Quad	Nastran-FSD	Static-AVL	Wing
3	Tri	Nastran-Traditional	Static-AVL	Wing
4	Quad	Nastran-Traditional	Static-AVL	Wing
5	Tri	HyperSizer	Static-AVL	Wing
6	Quad	HyperSizer	Static-AVL	Wing
7	Tri	Nastran-FSD	Aeroelastic-Nastran	Wing-Panel Tail
8	Quad	Nastran-FSD	Aeroelastic-Nastran	Wing-Panel Tail
9	Tri	Nastran-FSD	Aeroelastic-Nastran	Wing-Tail
10	Quad	Nastran-FSD	Aeroelastic-Nastran	Wing-Tail
11	Tri	Nastran-Traditional	Aeroelastic-Nastran	Wing-Panel Tail
12	Quad	Nastran-Traditional	Aeroelastic-Nastran	Wing-Panel Tail
13	Tri	HyperSizer	Aeroelastic-Nastran	Wing-Panel Tail
14	Quad	HyperSizer	Aeroelastic-Nastran	Wing-Panel Tail
15	Tri	HyperSizer	Aeroelastic-Nastran	Wing-Tail
16	Quad	HyperSizer	Aeroelastic-Nastran	Wing-Tail

ANOVA can be performed, as brought up in Section 3.5.6, to understand the effect of varying topology, optimization method, aerodynamics, and scope. This leads to Experiment 1.1.

Experiment 1.1 *Resolution and abstraction are proposed in the literature as fundamental characteristics the drive model fidelity. By understanding the impact on the variability of the model responses with respect to scope, and comparing it to the impact of variability of other resolution and abstraction-driven settings, the inclusion of scope as a third aspect of this description can be examined and justified.*

	Option 1	Option 2
Topology	Tri	Quad
Optimization	Nastran	HyperSizer
Aerodynamics	Static	Elastic
Scope	Wing	Wing-Tail

Table 6.7: Modeling Choices for Showing Importance of Scope

Omitting the Nastran-Traditional optimization option, there are essentially two model-

ing choices for each of the four categories defined above, as shown in Table 6.7. When the model is aero-static, the horizontal tail is only represented in the aerodynamic model, which essentially assumes that it is fixed. As such, this could be considered equivalent in terms of scope to the aeroelastic wing-panel tail option, and as such is not listed for this case.

Given the 4 categories, at 2 settings each, 16 different combinations can be enumerated. However, the combination of Static/Wing-Tail is not in the defined model set listed above. This is due to the properties of the aero-static analysis: however the horizontal tail is defined, it will not impact the wing weight since the loads are not updated as the vehicle is sized. As such, the 4 options where this combination occurs can either be omitted or represented by the Static/Wing results, and both will be shown.

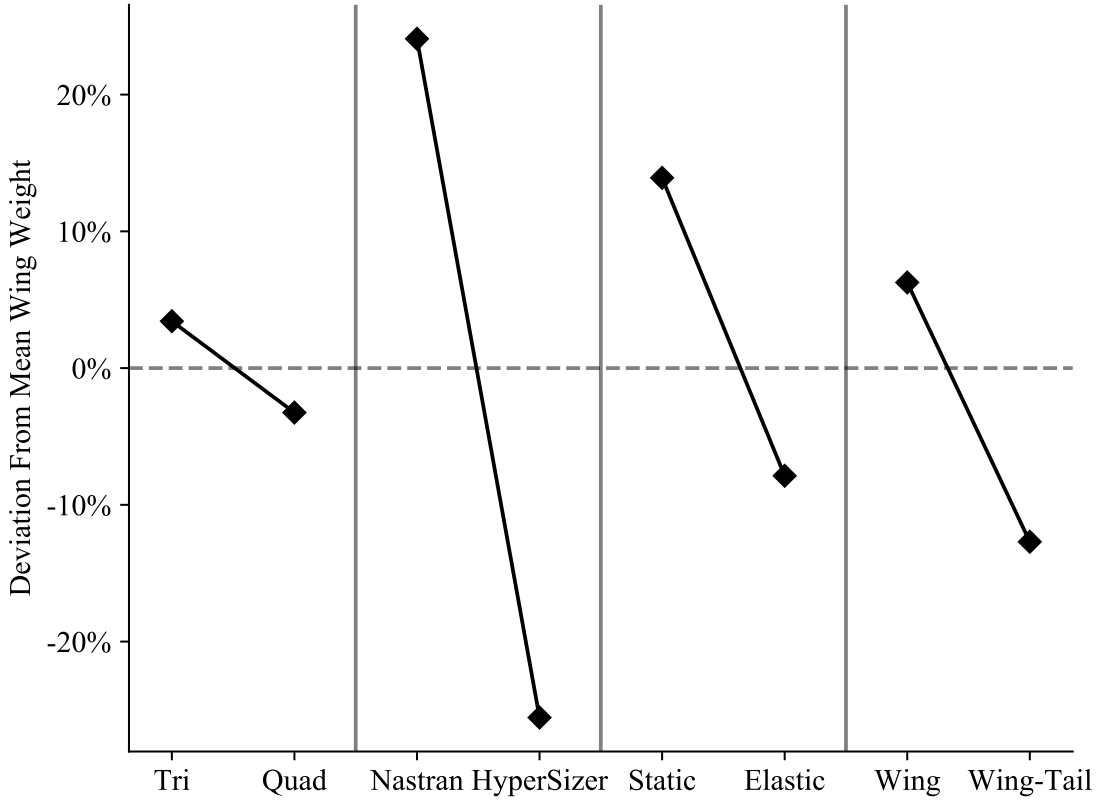


Figure 6.4: Main Effects Given 12 Cases

For each of the cases in the full-factorial combination of the options in Table 6.7, all

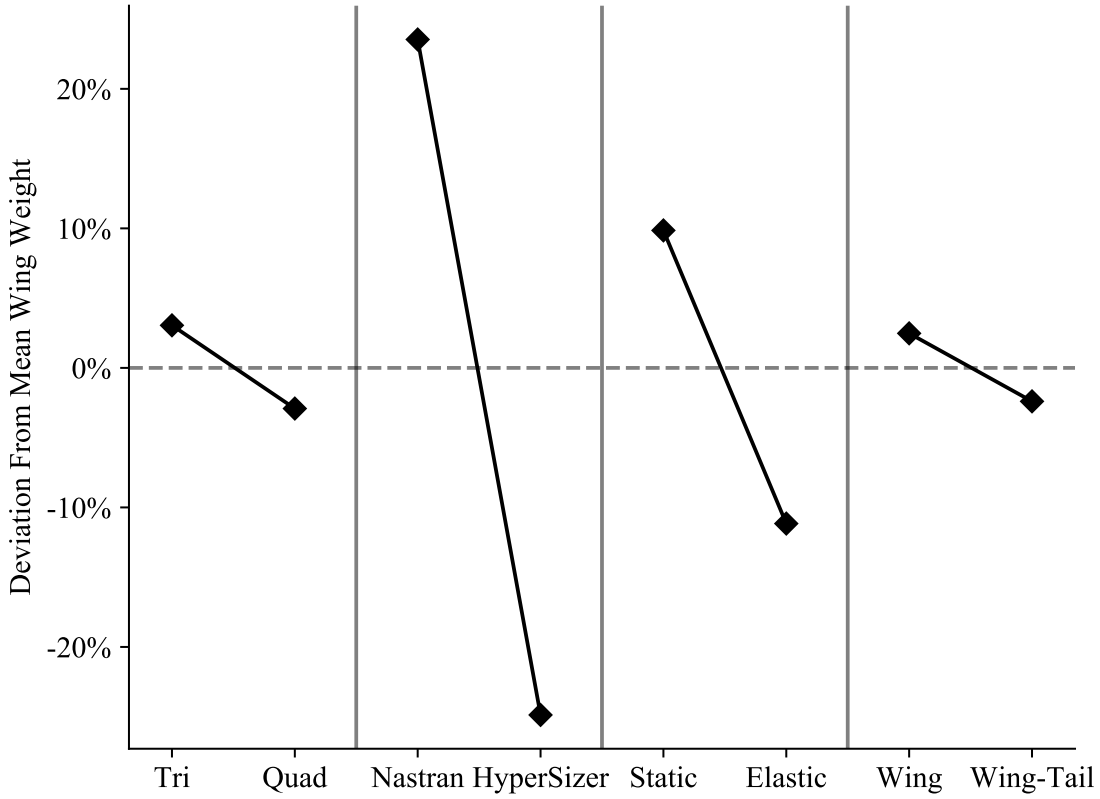


Figure 6.5: Main Effects, Using Static/Wing Results Also For Static/Wing-Tail

of the wing weight values are averaged together, and then averaged again to represent the variability in prediction between each option for the four categories. These results, omitting the four Static/Wing-Tail cases, are shown in Figure 6.4. When the Static/Wing results are used to stand in for the four excluded options, the main effects are shown in Figure 6.5.

While the impact of of scope is much more significant in Figure 6.4, there is still a distinct trend in Figure 6.5, showing that a change in scope causes a change in the variability of the estimated wing weight. For this type of plot, the impact of a particular effect is also dependent on the interaction of that effect and the others. However, the importance of the interactions with scope in this case have already been discussed. The tail cannot impact the sizing of the wing if the tail is not represented in the model. However, when the abstraction of the model is defined as aero-static, even if the tail is represented in the model, sizing the horizontal tail will have no effect on the estimation of sized wing weight.

This is similar to the representation of stiffeners in HyperSizer as defined in a previous section. In that case, as the resolution was increased, the interaction with abstraction did not adequately account for the change, so the overall result is less accurate. In this case, a change in scope can be an improvement, but, as before, only if the abstraction allows it to be. This leads to Conclusion 1.1.

Conclusion 1.1 *The consideration of scope as a fundamental characteristic of fidelity is justified not only by the variability in prediction that is causes, but in the way that it interacts with resolution and abstraction. As such, when decsribing model fidelity, resolution, abstraction, and scope should all be taken into account.*

6.4 Step 3: Descriptive Fidelity Assessment

Now that a set of sixteen models has been defined, the next step is to perform an initial fidelity assessment with respect to resolution, abstraction, and scope. The descriptive ordering for those three attributes are shown in Table 6.8.

Table 6.8: Initial Model Set 3 Assessment by Ordering

Fidelity Types	Model Ordered Groups
Resolution (1)	[Tri/FSD/Static/Wing, Quad/FSD/Static/Wing, Tri/Trad/Static/Wing, Quad/Trad/Static/Wing, Tri/FSD/Elastic/Wing-Panel Tail, Quad/FSD/Elastic/Wing-Panel Tail, Tri/Trad/Elastic/Wing-Panel Tail, Quad/Trad/Elastic/Wing-Panel Tail], [Tri/FSD/Elastic/Wing-Tail, Quad/FSD/Elastic/Wing-Tail], [Tri/HS/Static/Wing, Quad/HS/Static/Wing, Tri/HS/Elastic/Wing-Panel Tail, Quad/HS/Elastic/Wing-Panel Tail], [Tri/HS/Elastic/Wing-Tail, Quad/HS/Elastic/Wing-Tail]

Continued on next page

Table 6.8: Initial Model Set 3 Assessment by Ordering

Fidelity Types	Model Ordered Groups
Resolution (2)	[Tri/FSD/Static/Wing, Quad/FSD/Static/Wing, Tri/Trad/Static/Wing, Quad/Trad/Static/Wing], [Tri/FSD/Elastic/Wing-Panel Tail, Quad/FSD/Elastic/Wing-Panel Tail, Tri/Trad/Elastic/Wing-Panel Tail, Quad/Trad/Elastic/Wing-Panel Tail], [Tri/FSD/Elastic/Wing-Tail, Quad/FSD/Elastic/Wing-Tail], [Tri/HS/Static/Wing, Quad/HS/Static/Wing], [Tri/HS/Elastic/Wing-Panel Tail, Quad/HS/Elastic/Wing-Panel Tail], [Tri/HS/Elastic/Wing-Tail, Quad/HS/Elastic/Wing-Tail]
Abstraction (1)	[Tri/FSD/Static/Wing], [Quad/FSD/Static/Wing], [Tri/Trad/Static/Wing], [Quad/Trad/Static/Wing] [Tri/HS/Static/Wing], [Quad/HS/Static/Wing], [Tri/FSD/Elastic/Wing-Panel Tail], [Quad/FSD/Elastic/Wing-Panel Tail], [Tri/FSD/Elastic/Wing-Tail], [Quad/FSD/Elastic/Wing-Tail], [Tri/Trad/Elastic/Wing-Panel Tail], [Quad/Trad/Elastic/Wing-Panel Tail], [Tri/HS/Elastic/Wing-Panel Tail], [Quad/HS/Elastic/Wing-Panel Tail], [Tri/HS/Elastic/Wing-Tail], [Quad/HS/Elastic/Wing-Tail]
Abstraction (2)	[Tri/FSD/Static/Wing], [Quad/FSD/Static/Wing], [Tri/Trad/Static/Wing], [Quad/Trad/Static/Wing] [Tri/HS/Static/Wing], [Quad/HS/Static/Wing], [Tri/FSD/Elastic/Wing-Panel Tail], [Quad/FSD/Elastic/Wing-Panel Tail], [Tri/Trad/Elastic/Wing-Panel Tail], [Quad/Trad/Elastic/Wing-Panel Tail], [Tri/HS/Elastic/Wing-Panel Tail], [Quad/HS/Elastic/Wing-Panel Tail],

Continued on next page

Table 6.8: Initial Model Set 3 Assessment by Ordering

Fidelity Types	Model Ordered Groups
	[Tri/FSD/Elastic/Wing-Tail], [Quad/FSD/Elastic/Wing-Tail], [Tri/HS/Elastic/Wing-Tail], [Quad/HS/Elastic/Wing-Tail]
Scope (1)	[Tri/FSD/Static/Wing, Quad/FSD/Static/Wing, Tri/Trad/Static/Wing, Quad/Trad/Static/Wing, Tri/HS/Static/Wing, Quad/HS/Static/Wing], [Tri/FSD/Elastic/Wing-Panel Tail, Quad/FSD/Elastic/Wing-Panel Tail, Tri/Trad/Elastic/Wing-Panel Tail, Quad/Trad/Elastic/Wing-Panel Tail, Tri/HS/Elastic/Wing-Panel Tail, Quad/HS/Elastic/Wing-Panel Tail], [Tri/FSD/Elastic/Wing-Tail, Quad/FSD/Elastic/Wing-Tail, Tri/HS/Elastic/Wing-Tail, Quad/HS/Elastic/Wing-Tail]
Scope (2)	[Tri/FSD/Static/Wing, Quad/FSD/Static/Wing, Tri/Trad/Static/Wing, Quad/Trad/Static/Wing, Tri/HS/Static/Wing, Quad/HS/Static/Wing, Tri/FSD/Elastic/Wing-Panel Tail, Quad/FSD/Elastic/Wing-Panel Tail, Tri/Trad/Elastic/Wing-Panel Tail, Quad/Trad/Elastic/Wing-Panel Tail, Tri/HS/Elastic/Wing-Panel Tail, Quad/HS/Elastic/Wing-Panel Tail], [Tri/FSD/Elastic/Wing-Tail, Quad/FSD/Elastic/Wing-Tail, Tri/HS/Elastic/Wing-Tail, Quad/HS/Elastic/Wing-Tail]

Due to the size of the model set, and relative complexity of the differences between modeling options, these can be somewhat difficult to decipher. However, it shows just how difficult the previous method would be, simply assessing the fidelity of the model directly with specific values. This is a much more streamlined and digestible process, ordering the models with respect to easier to understand criteria.

While these orders are used for this work, they do not necessarily represent the only possible order that could be given. For example, all of the HyperSizer models are listed as higher resolution than all of the other models, since they represent more detail about the stiffened panels. However, someone might say that the horizontal tail detail present in the FSD/Elastic/Wing-Tail models should place it above the HS/Static/Wing models. A similar argument could be made for whether the HyperSizer failure modes mean that the HS/Static models have less abstraction than the non-HyperSizer/Elastic models. This again shows the power of this method, since, if both orders are believable enough, they could both be included.

The power of using orders based on resolution, abstraction, and scope should not be underestimated. However, the arguments made in the previous paragraph emphasize just how complex and case-specific fidelity assessment is. In certain cases, particularly for resolution and scope, it could be possible to assign hard numbers to the orders instead of just an order.

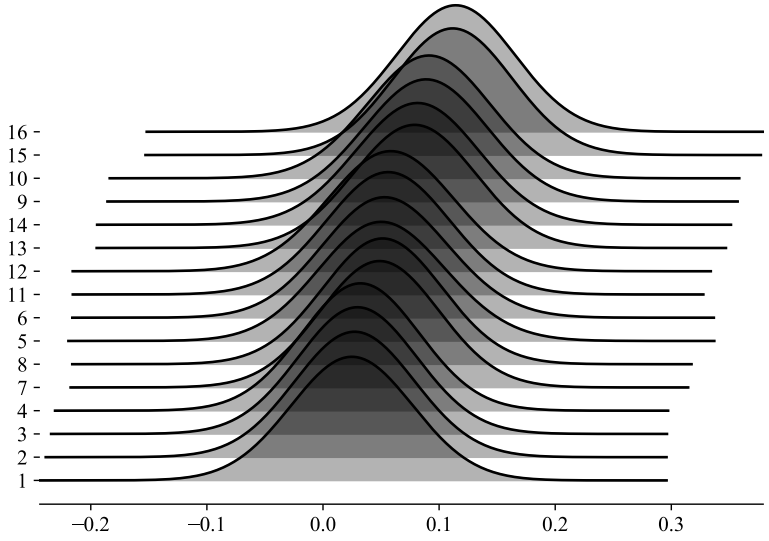
For example, if two very similar aircraft finite element models are being compared, but one is of half of the aircraft, and the other the full aircraft, the scope of one is exactly twice the scope of the other. In that case, the descriptive fidelity assessment would estimate that the full vehicle model has a higher fidelity. However, if the vehicle is symmetric, as almost all are, and all of the flight conditions being assessed are symmetric, then all of the shared responses that can be assessed between the two models should be identical.

Specifically to this model set, the smeared stiffness approach and failure modes in HyperSizer raise the resolution and lower the abstraction a great deal. Determining where the resolution of an unstiffened aeroelastic model fits in is complicated because, as mentioned previously, the aeroelastic effects will be greater as the aspect ratio increases. If the primary variables were changed, or even if the range of aspect ratios were adjusted, the true difference in abstraction would change. This is why fidelity assessment, even qualitatively, should be scenario-driven, and also why it is important to leverage whatever data is

available to refine to quality of the evaluation.

6.4.1 Model Probabilities

Using the orders in Table 6.8, the combined fidelity density estimates can be found, as shown in Figure 6.6.



(a) KDE

Median	Model ID
0.0250	1
0.0274	2
0.0299	3
0.0323	4
0.0488	7
0.0513	8
0.0516	5
0.0541	6
0.0561	11
0.0585	12
0.0785	13
0.0810	14
0.0883	9
0.0908	10
0.1117	15
0.1143	16

(b) Median Values

Figure 6.6: Aircraft Model Set Descriptive KDE

The density estimates show that there is a decent amount of agreement between resolution, abstraction, and scope since the variance of the densities all appear similar and the peaks generally follow a linear trend. Looking further at the values of the medians, though, does show how the bottom four models, the non-HyperSizer aero-static models, seem to be in a somewhat separate group. Similarly, the aeroelastic, HyperSizer, full tail models are mildly separated and are the top two options, which is expected as it is clear they have the highest resolution, lowest abstraction, and largest scope of any models.

The set of model probabilities based on these orders is shown in 6.7. As assessed

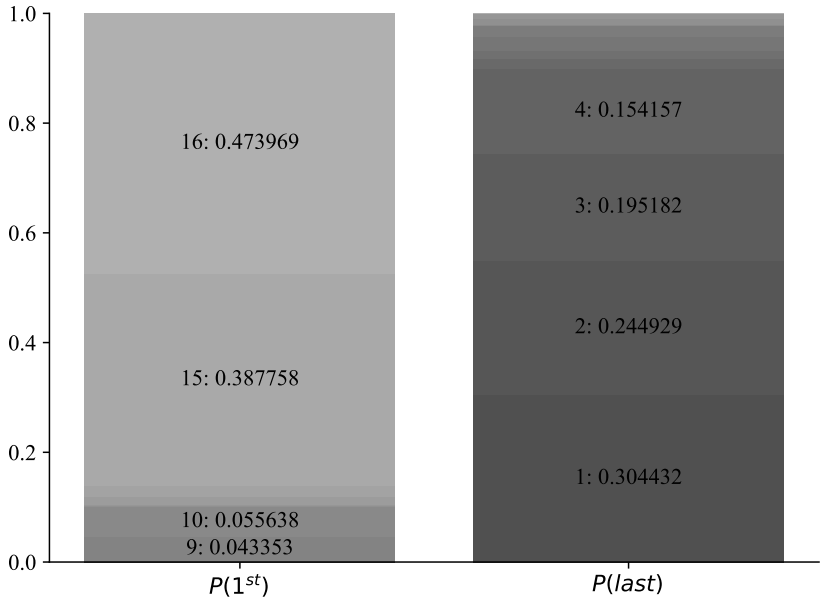


Figure 6.7: Model Set 3 Probabilities From Qualitative Assessment

from the distributions, the top two and bottom four options are clear, while the HyperSizer, aeroelastic, wing-panel tail models are also decent. This method of estimating fidelity based on descriptive assessment provides an easy platform for visually confirming the assumed relative relationships and calculating a set of model probabilities that can be used to select a model. However, because of the agreement between the aspects of fidelity, the model probabilities may be over-confident in how much more effective the top two models are compared to the rest of the set. More information could be gleaned by looking at the intermediate relative probabilities (e.g. $P(2^{nd})$, $P(3^{rd})$, ..., $P(penultimate)$) if desired, but since model data is available, this will be skipped for now in favor of acomparative data assessment to be performed in a following section.

6.5 Step 4: Fidelity Estimation Via Comparative Data Assessment

From Hypothesis 1.3, correlation and error metrics are presumed to aid in the understanding of a model set, since models that take different approaches to make the same estimation should agree with each other if they are truly appropriate for tackling the defined problem.

As a model set is being defined, those with expertise are hypothesizing that models have the appropriate representation, or are incorporating the correct phenomenology. As model data becomes available, outliers in comparative data assessment either require troubleshooting and debugging, or the expert's presumptions were incorrect. If the shape of the responses surfaces differ drastically, one of the models is accounting for something that another is not, which can be justification for exclusion from further consideration.

6.5.1 Model Results

While model set two was efficient enough to evaluate on a single machine, this set was distributed to three similar machines. The cases would have run faster if a distributed computing cluster could have been used, but software licensing limitations made it simpler to evaluate these cases on desktop machines.

Unlike with the previous model set, some of the cases were unable to converge. Some of the issues present were software issues that could be overcome by restarting the case, but other, due to the deterministic nature of the analysis, would not finish even with repeated evaluation. This increases the difficulty of comparative data analysis, and must be adressed prior to calculation of regrssion metric scores.

Similarly to the linear static deflection of the I-beam model set, a secondary response was extracted from this model set for the purpose of verification and begin to understanding their behavior. This is the wingtip deflection under cruise flight conditions, and are shown in Figure 6.8. As with the beam models of the previous chapter, the deflection must be extracted for the entire tip face. The translations in the x, y, and z directions are extracted from the HDF5 file for all of the wing tip rib nodes. These values are averaged together, and then the reported response is the magnitude of the deflection. The results in Figure 6.8 appear as expected:

- Deflection increases with aspect ratio
- The models are in relative agreement

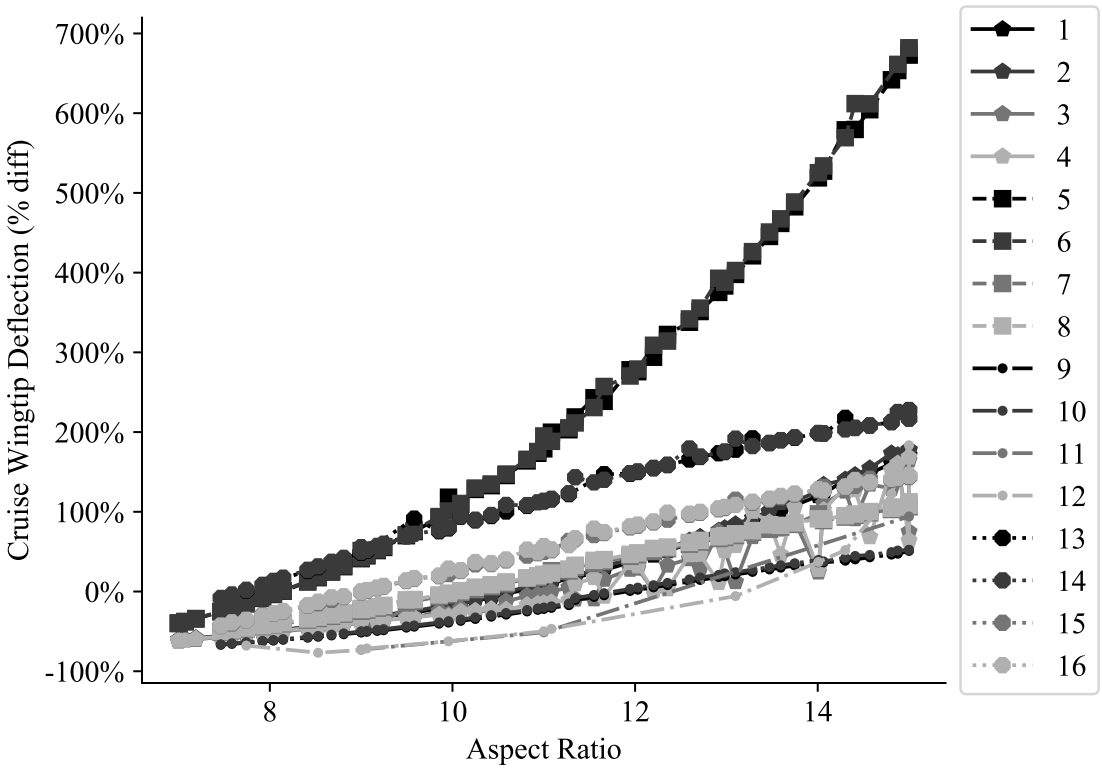


Figure 6.8: Cruise Wingtip Deflection (% difference from Quad/HS/Elastic/Wing-Tail, AR=9)

- Models 5, and 6 are aero-static but use HyperSizer, resulting in a flexible wing, but with no load updating, larger deflections

These results are good for verifying that the models seem to be working, but are not the most pertinent model response. The estimated wing weights are shown in 6.9. One thing becomes immediately obvious upon cursory data examination: two of the models return much higher weights than the rest of the set. This will be addressed more in the down-selection section.

Otherwise, the models are in relative agreement. The quick optimization FSD weights are more unstable, which is expected since this is not assumed to reach convergence. The HyperSizer results are the most well-behaved, and the HyperSizer aeroelastic results all share a similar trend and relative value.

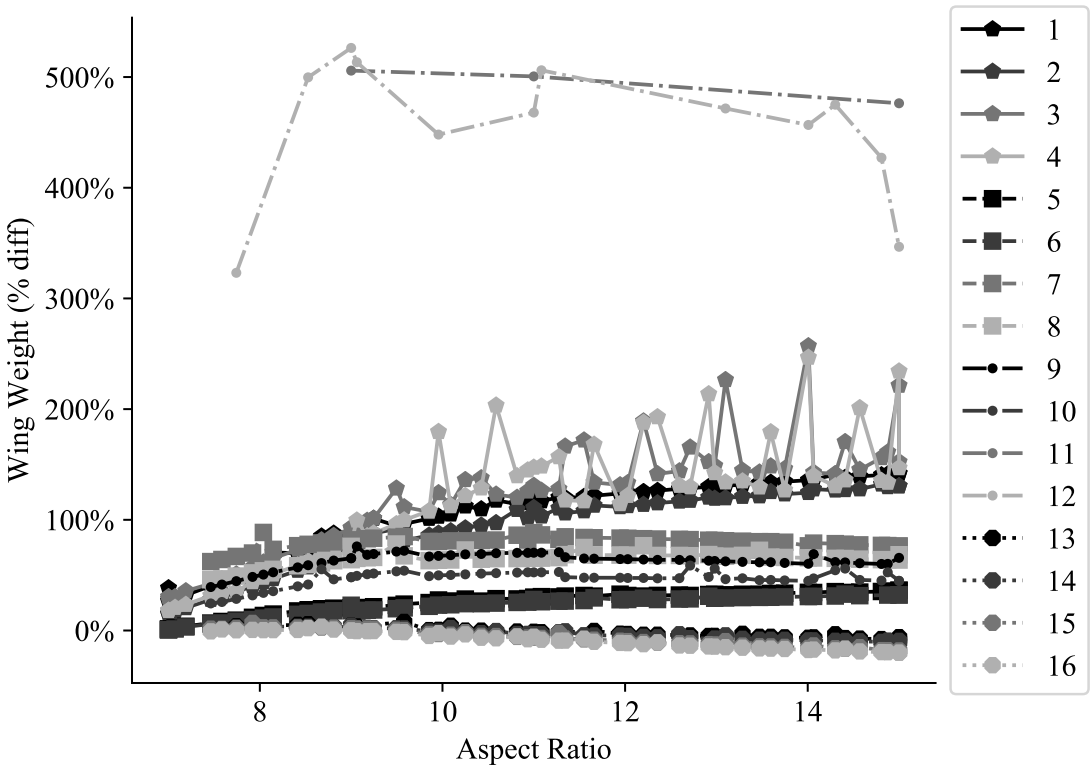


Figure 6.9: Wing Weight Estimates (% diff from baseline)

6.5.2 Data Alignment

As mentioned previously, the data points should be in alignment prior to calculation of correlation and error metrics. This was not a problem for the well-behaved models of the previous set, but with this model issues arise. Even though the same design of experiments was used, failed cases leave gaps in the data. Additionally, some of the models have a meshing issue that prevented them from running successfully for the low end of the range. This problem could potentially be alleviated with some work, so a model should not be written off on that basis.

The first step for data alignment between two models is finding the valid ranges. In this one-dimensional case, this is a relatively simple process. For models M_1 and M_2, if M_1 has data points in $[7, 15]$ but M_2 has only been tested in $[10, 17]$, the resulting shared range

is $[10, 15]$. For a multi-dimensional case, this process must be repeated in each dimension to find a shared n-dimensional box.

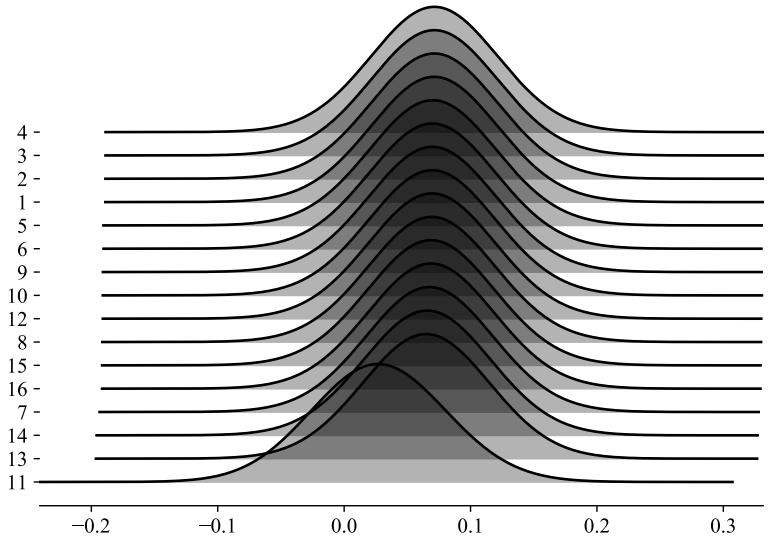
Within that range, the symmetric difference between the point locations for M_1 and M_2 are found. In the absence of additional information, linear interpolation is used to find missing points in each set. Specifically, the *griddata* method in *SciPy* is used, which can interpolate n-dimensional data using convex hulls[124]. Using a simple linear interpolator is efficient and does not make too many assumptions about the data. Two of the common situations it is attempting to alleviate is a hole in the data due to a failed case, and minor misalignment of evaluation points, both of which it should handle well. Specifically, between most of the models in this set, only one or two interpolations are needed, if that. The primary exceptions are models 11 and 12, since only three points were successfully evaluated by model 11.

6.5.3 Correlation and Error Scoring

Once a shared set of data has been found, the R^2 and $RMSE$ are calculated between each set. The arrays are then normalized according to the previously described process to arrive at two new sets of fidelity scores. Figure 6.10 shows the density estimates based only on the R^2 and RMSE scores. Both sets of scores further emphasize that something is wrong with model 11, and the RMSE scores include model 12 as well. This is important, because for a multi-dimensional problem, the problem might not be as obvious as in Figure 6.9.

Clearly the error between those two models and the rest of the set is large, but R^2 is not as clear about model 12. This can be explained by looking at the trend of the three points that were evaluated for model 12. While their magnitude is much different, the slight downward trend is very similar to many of the other result sets. Because of this the R^2 between model 12 and some of the other values is relatively high.

The results can be examined further in Figure 6.12, which shows the distributions based on combination of the R^2 and RMSE scores. Models 11 and 12, the Trad/Elastic/Wing-



(a) KDE

Median	Model ID
0.0276	11
0.0648	13
0.0657	14
0.0675	7
0.0687	16
0.0687	15
0.0693	8
0.0694	12
0.0697	10
0.0698	9
0.0698	6
0.0701	5
0.0713	1
0.0713	2
0.0713	3
0.0714	4

(b) Median Values

Figure 6.10: Sorted Density Estimates Based on R^2 Scores for Aircraft Use Case

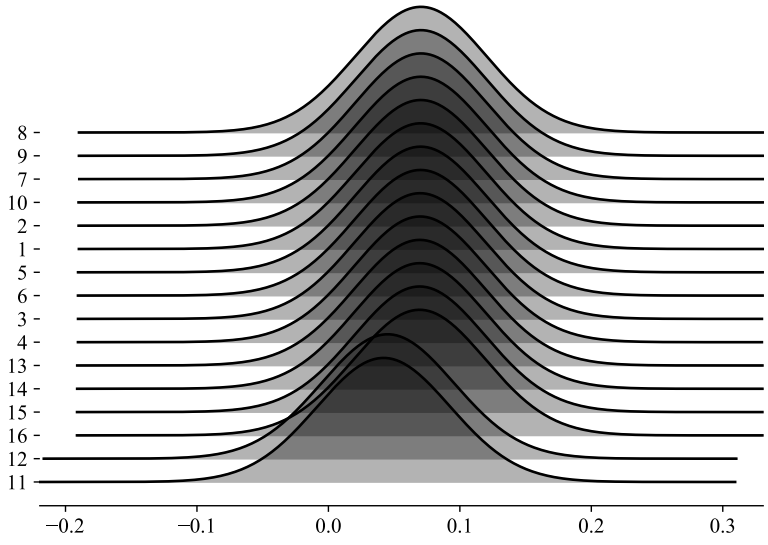
Panel Tail models, are both at the bottom of the list.

Interestingly, since the HyperSizer models have more of a downward trend than many of the other models as aspect ratio increases, they only rank in the middle of the pack based on the correlation assessment. However, this implications of this are minimal as most of the distributions appear similar aside from the aforementioned exceptions.

Now that the correlation and error metrics have been calculated, this is the opportunity for potential initial down-selection followed by the adjusted assessment of fidelity.

6.5.4 Step 4.1: Initial Down-Selection

Unlike for model set two, none of the pairwise model comparisons have both a high-enough R^2 and low-enough $RMSE$ to justify removal from the set. However, as mentioned in a previous section, the two aeroelastic, Nastran standard design cycle optimization models proved difficult to analyze. Many of the cases failed, and as you can see in Figure 6.9, the



(a) KDE

Median	Model ID
0.0422	11
0.0452	12
0.0693	16
0.0694	15
0.0695	14
0.0695	13
0.0697	4
0.0697	3
0.0701	6
0.0701	5
0.0701	1
0.0703	2
0.0703	10
0.0704	7
0.0704	9
0.0704	8

(b) Median Values

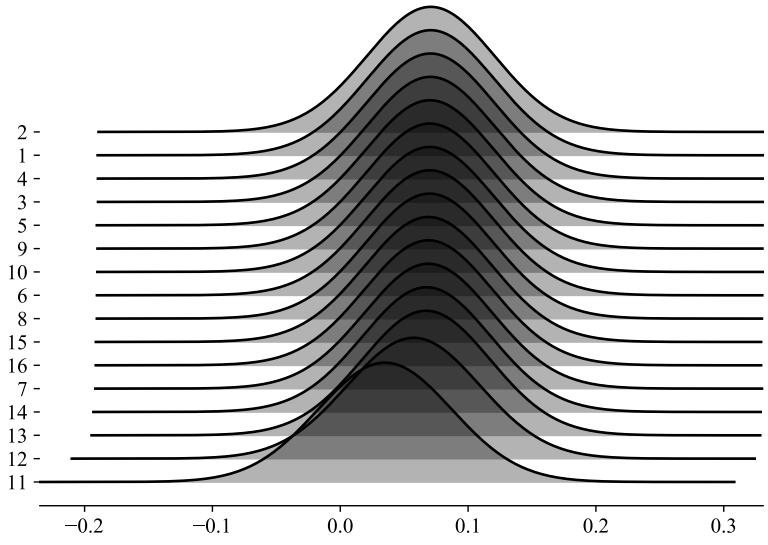
Figure 6.11: Sorted Density Estimates Based on $RMSE$ Scores for Aircraft Use Case

responses that did converge were completely out of the range of the other results.

As mentioned above, model 11, the Tri/Trad/Elastic/Wing-Panel Tail model provided responses that have a similar trend to the HyperSizer responses. This does not, however, justify its inclusion, since only three of the cases succeeded and the cases that did run took longer to evaluate than any others in the set.

In fact, many of the models timed out after two hours, when most of the HyperSizer models around 30 minutes. This is why the Trad/Elastic/Wing-Tail models were excluded entirely. The models are significantly less efficient in their current state than the HyperSizer models, which can provide higher resolution response, so they are not worth continued effort.

There is a possibility that some amount of tweaking to the optimization parameters in Nastran could correct these responses. However, all of the other models using Nastran optimization were behaved better with the same set of optimization settings. As such, the



(a) KDE

Median	Model ID
0.0350	11
0.0574	12
0.0672	13
0.0676	14
0.0689	7
0.0690	16
0.0690	15
0.0698	8
0.0700	6
0.0700	10
0.0701	9
0.0701	5
0.0705	3
0.0705	4
0.0707	1
0.0708	2

(b) Median Values

Figure 6.12: Sorted Density Estimates Based on Combined R^2 and $RMSE$ Scores for Aircraft Use Case

aeroelastic full Nastran optimization models are removed from the set for the remainder of this analysis. The resulting predicted weights are shown in Figure 6.13, and the new set is listed in Table 6.9. As with the I-beam model, the descriptive fidelity orders are adjusted simply by omitting these models, and are otherwise unchanged.

6.5.5 Adjusted Model Probabilities

After the initial processing of data and down-selection, the correlation scores are combined with the expert-elicited fidelity scores to generate adjusted distributions, as shown in Figure 6.14.

A few observations can be made based on the adjusted density estimates and the corresponding median values in the table in Figure 6.14. The two HS/Elastic/Wing-Tail models still have the highest estimated fidelity by a noticeable margin.

However, interestingly, the rest of the models have sorted themselves into pairs by

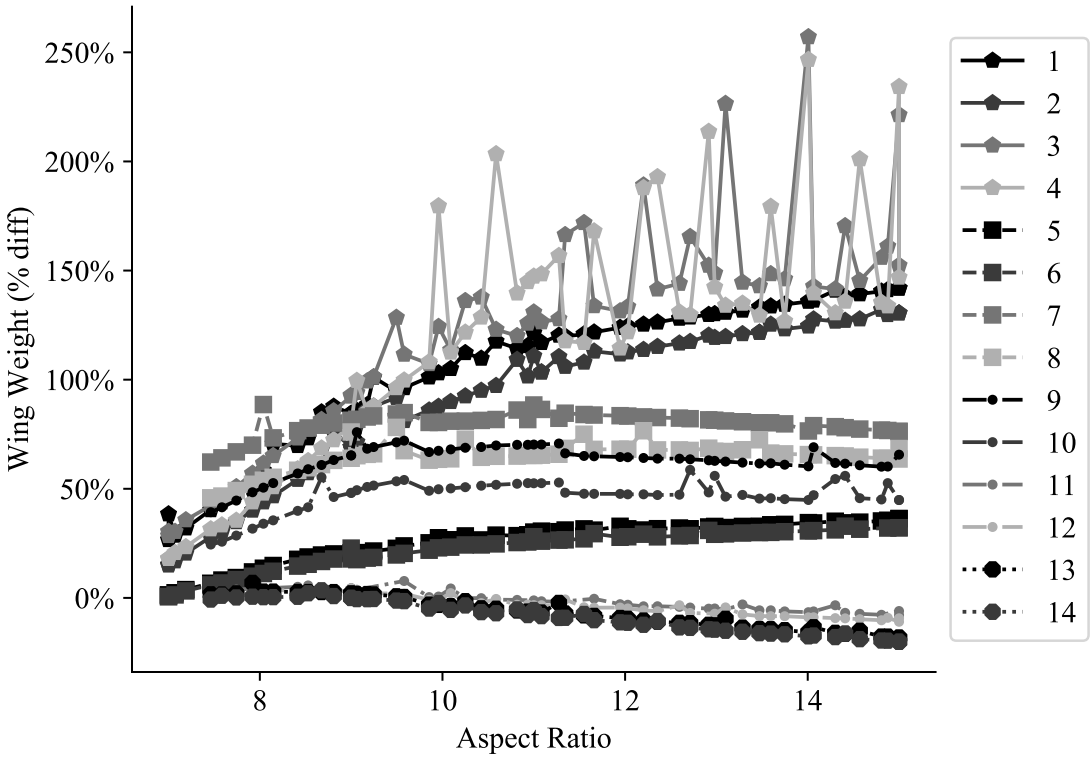


Figure 6.13: Wing Weights for 14 Aircraft Models (% diff from baseline)

topology. There is not a clear enough difference between all of the groups for to provide too much confidence for this based only on the fidelity estimate. However, knowing that the tri elements should not be as accurate as the quads agrees with the fact that for each pair, both tri models are lower than the corresponding set of quad representations.

Another observation is that the four aero-static models using the Nastran optimizer remain in a group of their own at the bottom, though the group is closer than before once correlations are included. Using the same method as before, a new set of adjusted model probabilities is calculated, as shown in Figure 6.15. The $P(last)_{adj}$ values again show that, while they account for less of the probability than before, models 1 through 4 are in a separate group, with the tri models scored lowest followed by the quad models. The pairing of probabilities based on topology is apparent throughout the adjusted probabilities.

The primary takeaway from the adjusted model probabilities, aside from helping to

Table 6.9: Model Set 3 After Initial Down-Selection

ID	Topology	Optimization	Aerodynamics	Scope
1	Tri	Nastran-FSD	Static-AVL	Wing
2	Quad	Nastran-FSD	Static-AVL	Wing
3	Tri	Nastran-Traditional	Static-AVL	Wing
4	Quad	Nastran-Traditional	Static-AVL	Wing
5	Tri	HyperSizer	Static-AVL	Wing
6	Quad	HyperSizer	Static-AVL	Wing
7	Tri	Nastran-FSD	Aeroelastic-Nastran	Wing-Panel Tail
8	Quad	Nastran-FSD	Aeroelastic-Nastran	Wing-Panel Tail
9	Tri	Nastran-FSD	Aeroelastic-Nastran	Wing-Tail
10	Quad	Nastran-FSD	Aeroelastic-Nastran	Wing-Tail
11	Tri	HyperSizer	Aeroelastic-Nastran	Wing-Panel Tail
12	Quad	HyperSizer	Aeroelastic-Nastran	Wing-Panel Tail
13	Tri	HyperSizer	Aeroelastic-Nastran	Wing-Tail
14	Quad	HyperSizer	Aeroelastic-Nastran	Wing-Tail

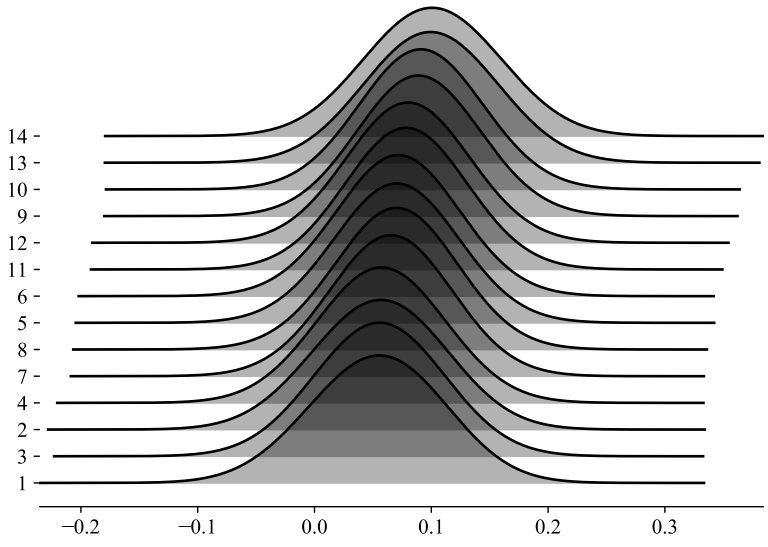
troubleshoot that the two models should be removed, is that most of the assumptions about the model set have been confirmed. The probabilities are more conservative, but the correlations have only made it more obvious that the behavior of the models is dependent on topology, which was not apparent based only on the descriptive assessment. This leads to Conclusion 1.3.

Conclusion 1.3 *Assessment of model agreement using the correlation and error metrics R^2 and $RMSE$ greatly increases the understanding of modeling options by classifying duplicates, identifying deficiencies caused by poor verification or a lack of representation, and updating qualitatively-derived probabilities of highest fidelity based on quantitative comparison.*

6.6 Step 5: Fidelity and Cost Scoring for Multi-Attribute Decision-Making

6.6.1 Step 5.1: Multifidelity Ranking

Now that the adjusted model probabilities have been calculated, the multifidelity rankings can be adjusted based on the correlations. As before, a truncated set of permutations is



(a) KDE

Median	Model ID
0.0540	1
0.0556	3
0.0559	2
0.0572	4
0.0652	7
0.0697	8
0.0702	5
0.0715	6
0.0785	11
0.0803	12
0.0893	9
0.0913	10
0.0997	13
0.1011	14

(b) Median Values

Figure 6.14: Correlation-Adjusted Fidelity Distributions for 14 Aircraft Models

assessed for efficiency since the model set still has fourteen models.

The models could be scored with respect to fidelity on its own, but, since cost data is available, more of the options can be analyzed if done with respect to the multi-attribute problem. This alleviates the memory requirement, though waiting on all of the possible combinations of 14 models is still very time-consuming. As discussed before, scoring all possible permutations for a model set of this size is also unnecessary, as it includes combinations that are unreasonable from an implementation standpoint.

6.6.2 Step 5.2: Cost/Efficiency Scoring

In addition to the estimated wing weight and cruise wingtip deflection, the times required for model creation, evaluation, and post-processing were recorded for each case. The analysis time for a case that is purely run through Nastran is easy to find in a reliable manner, since the computational time is reported separately from elapsed real time, and can be read from a Nastran-generated log file. The other times, including the analysis time when HyperSizer is involved, must be tracked simply by finding the difference between two Python

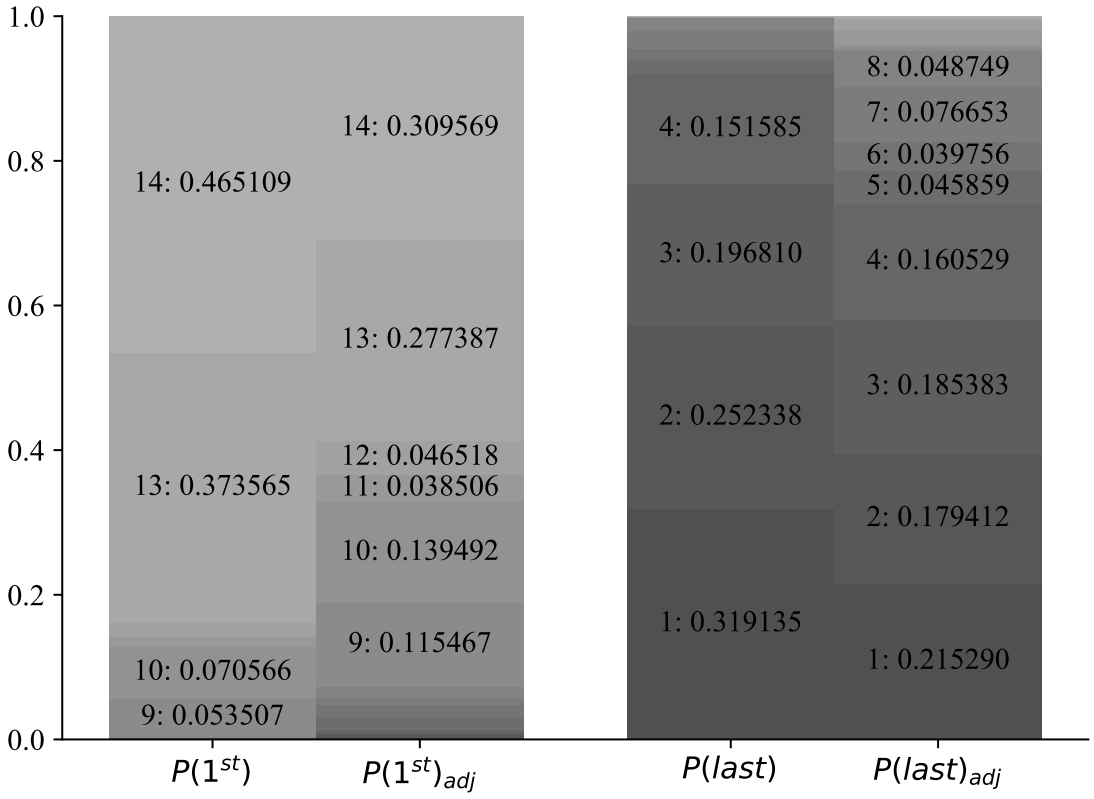


Figure 6.15: Correlation-Adjusted Model Probabilities for 14 Aircraft Models

wall time instances.

Instead of a single desktop, as with the previous model set, the evaluation of this model set was spread amongst three desktops. This sped up the process via parallelization, and three machines with similar specifications were selected to avoid bias. After removing outliers caused by outside strain on the system's resources, the generation and post-processing times were similar for all models, so the comparison of computation burden can be narrowed down to the analysis or evaluation time.

This was not unexpected since all of the models are shell Nastran models, were generated using RADE, and post-processed using similar code. Additionally, it could be noted that the generation of models could be made much more efficient by only processing each unique geometry once. However, the capability to save the current state and come back to it more than once is currently under development, and was not available at the time the

model evaluations were performed.

From Figure 6.16, with the exception of model 8, which will be discussed more below, the highest cost models are clearly 5, 6, and 11-14, or, those that use HyperSizer. Both in terms of the variables needed to represent stiffeners, and failure modes, using HyperSizer represents a dramatic increase in the number of effects taken into account. In addition, since the solution has to iterate between HyperSizer and Nastran, the recorded time must simply be the recorded wall time, which is at least part of what leads to the increased variability in the distributions of cost samples.

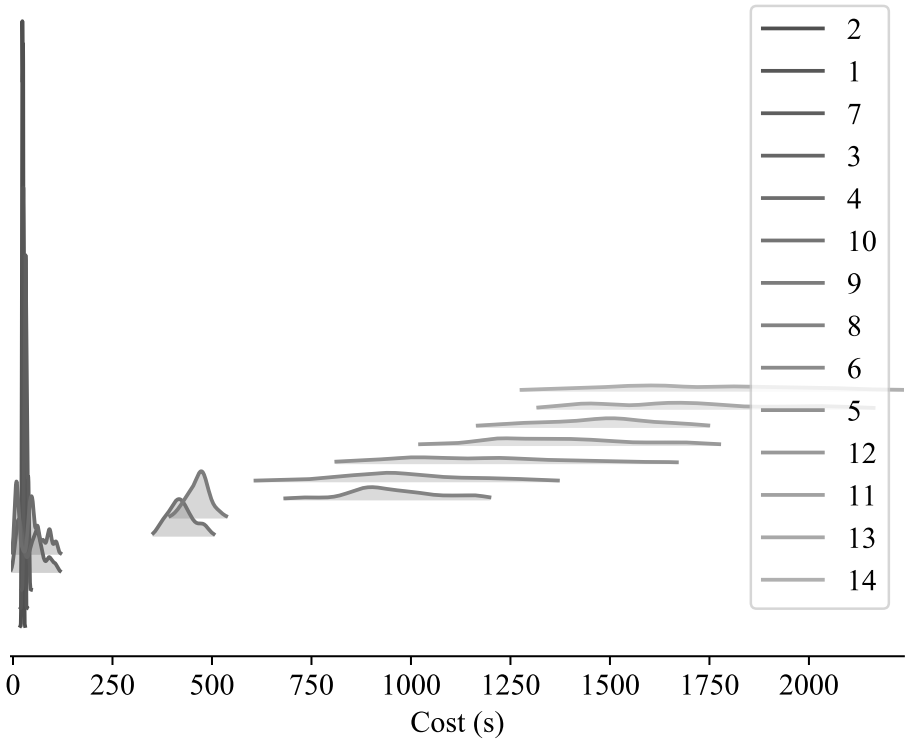
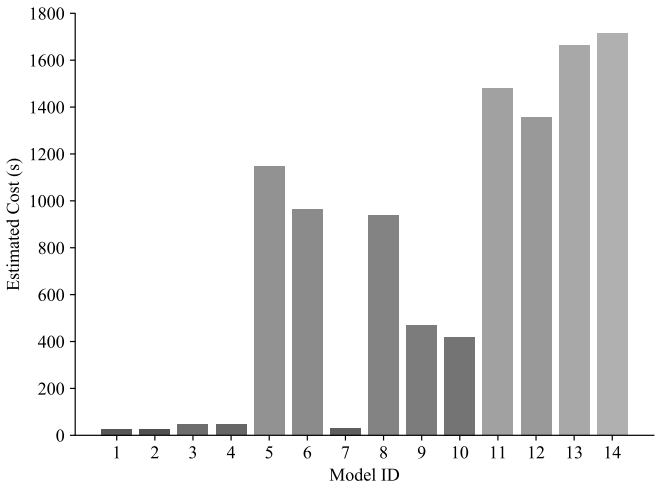


Figure 6.16: Cost Distributions with Outliers Removed for Down-Selected Set of 14 Aircraft Models

Figure 6.17 shows that while visualization of the distributions can provide some insight, it is difficult to easily distinguish all of the relevant information. Based on the decision to use the data sets directly to evalute cost, as well as the tendency towards a normal dis-

tribution, the median of each set is used here as the estimate of model cost, as shown in Figure 6.17. In general, the median of a kernel density estimate could still be used, but will increase the evaluation time.



ID	Model Cost (seconds)
1	26.426000
2	24.522500
3	48.188000
4	46.550000
5	1143.502930
6	960.171099
7	33.306000
8	921.466000
9	472.651500
10	419.533000
11	1478.343477
12	1359.127512
13	1657.953840
14	1703.608598

Figure 6.17: Estimated Costs for Aircraft Model Set After Initial Down-Selection

For the most part, these results are what would be expected. The HyperSizer models should be the most expensive, the more complete Nastran optimization takes longer than the fully stressed design in most cases. However, the difference between models 7 and 8 is worth noting. The quad model, based on the data gathered, appears drastically more expensive than the tri model of the same type.

This could be investigated further in the future to see if some unstable interaction exists when performing an aeroelastic analysis with the simplistic fully stressed design routine. In fact, the reported times for both model 7 and 8 appear somewhat bi-modal, though without more data, it is hard to determine if this is a consequence of a region of the design space, optimization settings, or simply that the models were evaluated around the same time and the machine being used was performing some other task. Even if they were evaluated on different machines, a scheduled task could have started that tasked both computers around

the same time. Regardless, and since these models are likely not the highest fidelity in the set, for this work the cost estimates in the table in Figure 6.17 are what will be used moving forward.

Models 9 and 10, the FSD/Elastic/Wing-Tail models, show the increase in runtime that comes from representing the horizontal tail as a full flexible shell wingbox. The main consequence of this is in the number of design variables that must be added to incorporate the shell panels that make up that structure. When using HyperSizer, the difference in cost is less obvious due to the variability in the cost samples, but it can still be seen in that 13 and 14 are above 11 and 12.

Another takeaway of the cost estimation refers back to the adjusted model probabilities, where the models grouped by topology, with the tri models falling lower than their corresponding quad meshes. Tri meshes are easier to fit, but since the quad meshes seem to be performing well in this case, one of the only justifiable reasons for using a tri mesh is if it allows for a more efficient intermediate assessment.

However, the estimated costs are mostly independent of topology. Because of this, the argument could be made to pick one of the topologies, likely the quads, and remove the other models. However, for this work, and since there is nothing that has shown to be inherently wrong with the tri models, the same model set will continue to be used. Using the estimated costs and cost ratios in the model set, single and multi-model efficiency scores can be generated, which will be discussed in terms of the multi-attribute decision-making process in the section.

6.6.3 Step 5.3: Multi-Attribute Decision-Making

The permutations of the down-selected aircraft model set are scored in terms of both fidelity and efficiency. For more informationon the multifidelity scoring, refer back to Algorithm 2 and the related section. For information on multi-model efficiency scoring, see Equations 5.11 and 5.12 the surrouding section. These methods were tested out using a notional and

I-beam FEM model set in Section 5.8.

The total number of permutations for the down-selected aircraft model set is 236,975,164,804, which would be the case for any set of 14 models. The individual models are scored in terms of fidelity and efficiency and shown relative to the single-model set of non-dominated models, or Pareto front. From a development and implementation standpoint, smaller combinations of models are preferable, so the smaller permutations are evaluated first. When evaluating a set of ordered model collections, if the number of scores increases beyond one million, the method is run to comvert that full list to a set of non-dominated orders. This dramatically decreases the memory requirements for ranking, as the number of non-dominated multifidelity options is typically below 25 for the cases that have been tested.

For the aircraft model set, the cost ratio between the slowest and fastest models is around 70, which is similar to Thunnissen's comment referenced in Chapter 4 that high fidelity models can easily be more than 100 times more costly[2]. This case is also similar to the notional case with a geometric progression of cost in Figure 5.26b. In this case, all of the models are even of the level to be applied to a similar point in the design process, so such a difference in cost should have an effect on decision-making.

The Pareto fronts based on the descriptive assessment of fidelity is shown in Figure 6.18. Of the 236 billion permutations, the scoring stopped after evaluating the first 24,726,075 and finding 24 non-dominanted points. All of the 2-7 length orders were scored, as well as over 6 million of the 8-model ordered sets.

As shown in Figure 6.18, 6 of the single model options are in the dominated region, though model 9 is close to model 10. Since the top two fidelity models are very similar in cost, the combination of them, while it maintains a high fidelity score, has an efficiency score much higher than any other option shown. Based on this assessment few of the multifidelity options, only three of the multifidelity options come close to the desired upper-left corner: 14-13-10, 13-14-10, and 14-10. Model 10 takes less than less than 7 minutes to run, as opposed to the over 28 minutes per evaluation of model 14.

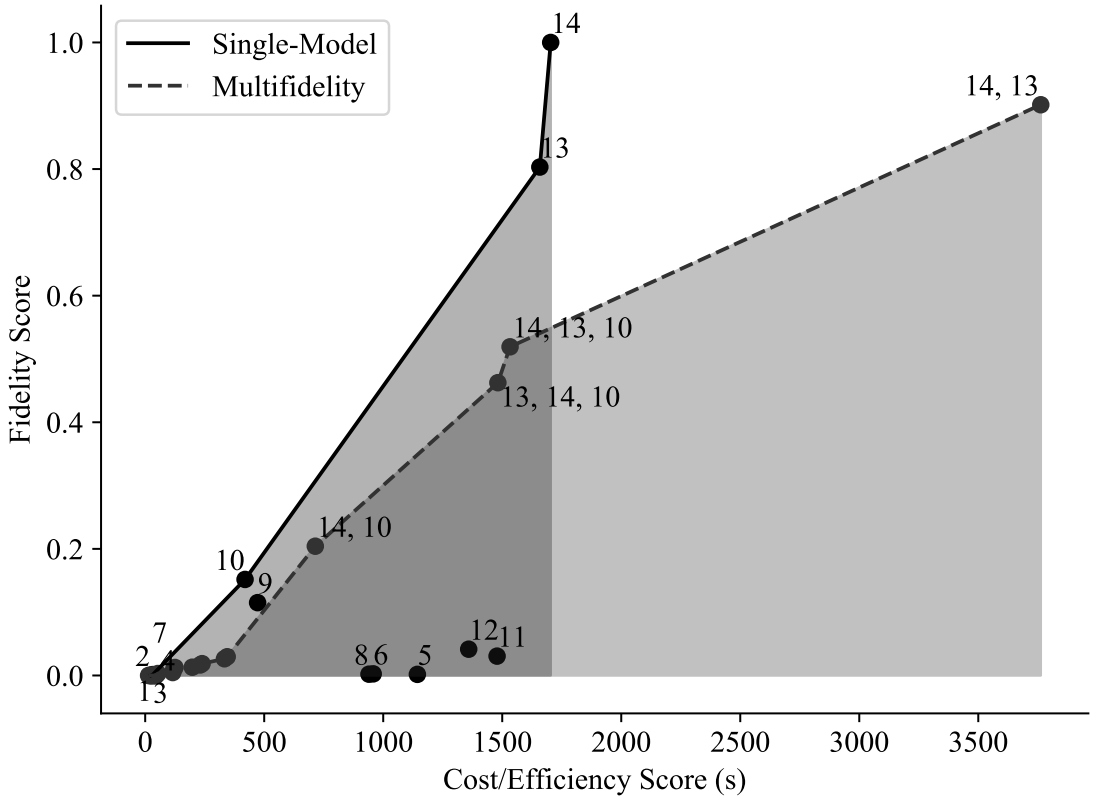


Figure 6.18: Single and Non-Dominated Multi-Model Ordered Combinations for Aircraft Model Set

After adjusting the fidelity probabilities based on comparative data analysis, the updated Pareto fronts are shown in Figure 6.19. The routine stopped after evaluating 22,726,073 permutations. Most of the overall comments prior to correlation scoring remain true:

- The difference in cost and fidelity does not lead to multifidelity options that fall to the left of the single model Pareto front
- Permutation 14-13 is the highest fidelity multifidelity option, but at a large cost increase
- Options including models 14, 13, and 10 are potentially worth consideration despite not clearly falling to the left of the single model Pareto front

However, since the models agree with one another, including model 10 does not require as

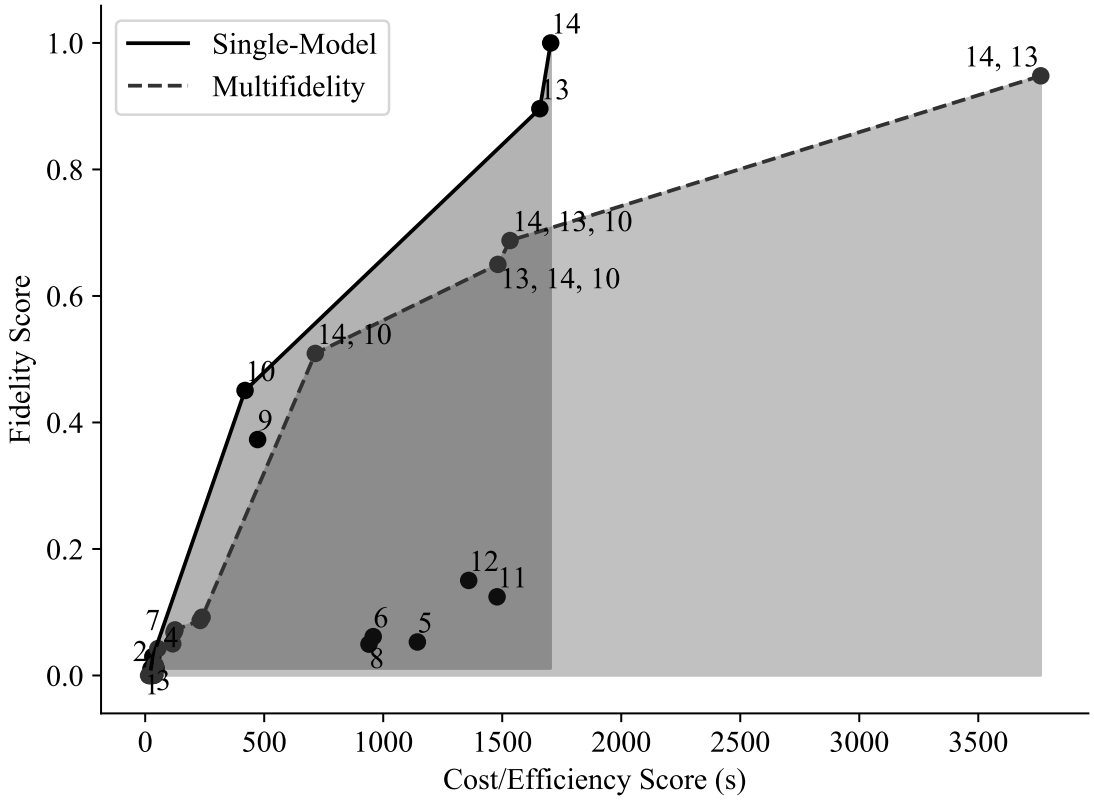


Figure 6.19: Correlation-Adjusted Single and Multi-Model Pareto Fronts for Aircraft Model Set

much of a detriment to the overall fidelity of the combination. In fact, the fidelity score for 14-10 is $\approx$0.5.

Even though the combination of models 14 and 10 does not fall clearly on the non-dominated side of the single models, it is still worth consideration for an important reason. Model 10 is the Quad/Nastran-FSD/Aeroelastic-Nastran/Wing-Tail model. This means that aeroelasticity is taken into account and the scope is at the highest level, but the stiffener representation and optimization routine are more of a rough approximation to find the correct order of magnitude for wing weight. Despite the limitations of the stiffener representation, it is one of the highest scoring models in the set in terms of fidelity.

All of the models in the set start their optimization routine from the same starting point: a uniform thickness across all of the panels. The final optimized structures taper the thick-

ness from root to tip in order to distribute the stress as evenly as possible. Because of the initial conditions, the costly HyperSizer failure mode and optimization routines must spend time trying to find the neighborhood of the appropriate thickness distribution before final optimization can be done.

Instead of using models 14 and 10 as two separate models combined using multifidelity regression, in this case the Pareto can be used to justify investigation of a different type of multifidelity approach: development of a two-phase optimization scheme. The Nastran-FSD model is generated first and runs more quickly to put the thicknesses in the right order of magnitude, tapering from root to tip of the wing. Then, if the HyperSizer model can be instantiated with the resulting thicknesses from model 10, HyperSizer will be in a much better starting position, reducing the overall runtime. In fact, since these models are similar in terms of topology, aerodynamics, and scope, the shell mesh can even be shared between them.

The impact of an initial, faster, optimization run is important in this case because the optimizer and solver are disconnected. HyperSizer optimizes the stiffnesses based on the current set of internal forces, then Nastran must be run to update the internal loads, and the process repeated until some measure of convergence is achieved. Because of this, if the initial starting point for HyperSizer is improved, the initial set of internal forces from Nastran is more realistic. As such, the first optimization of HyperSizer is much more effective.

While this type of multifidelity combination is not generally applicable, it is a technique that has been previously used in structural optimization. Additionally, it is discussed to point out how the multi-attribute decision-making process of generating Pareto fronts can be used. Importantly, even if the model builders do not want to go down the path of developing a combined approach, running model 14 by itself may still be too expensive at the point in the design process for which the decision-making process is being used. In such a case, the combinations 14-13-10 and 13-14-10 may also be too costly. In fact, if the effi-

ciency requirement is above 715 and below 1480 seconds, the combination of evaluations from models 14 and 10 still becomes the most favorable option in terms of fidelity. This assessment, as well as the assessment of the model sets in Chapter 5, lead to Conclusion 2.

Conclusion 2 *The evaluation of relative fidelity using the probability of highest fidelity available, as well as combined cost per evaluation of a single model or multi-model combination, is shown to be of great use in selecting the most appropriate model or models for continued development and evaluation given the current fidelity and efficiency requirements. This is proven through multi-attribute evaluation of fidelity and efficiency scores for a notional model set, I-beam FEM multifidelity set, and aircraft wing weight estimation trade study.*

6.6.4 Step 5.4: Selection of New Evaluation Points

A multi-attribute decision-making process does not have to be limited to fidelity based on a single response and estimated efficiency. If the fidelity based on a different model response was also important, a fidelity assessment could be generated using comparative analysis of those responses, and included to inform model down-selection.

When the data is available, model cost should always be included to prevent selection of a model or models that will only allow a small number of evaluations, as this will limit the understanding of the design space. Once an ordered combination is selected, the rules of thumb from the work of Toal should be used to determine how many cases to evaluate from each model.

This should be straightforward, as the pairwise correlations (R^2) have already been calculated in the assessment of fidelity and the cost ratios (C_r) have already been estimated in the process of determining the cost scores. From there, f_r can be selected through the order to determine the number of evaluations to perform for each model. From there, an appropriate design of experiments can be selected to attempt to gather as much additional information as possible in an efficient manner.

However, evaluation of selected models is not just dependent on the number of evaluations to be performed, but the location of those points in the design space. The selection of evaluation points is related to the field of Design of Experiments already referenced earlier in this work. If no model data has been generated, standard structured or space-filling designs can be used to efficiently cover the variable ranges as efficiently as possible. Examples of standard designs include those in the *pyDOE2* Python package used in this work: factorial designs, Plackett-Burman, Box-Behnken, Central Composite, Latin-Hypercube, etc.

In the case that some amount of model data has been generated, it should first be used for comparative data analysis in step 4. After a model or models have been selected, however, a standard design of experiments may no longer be as efficient. If a minimal amount of edge cases have been evaluated, a space-filling design may still be appropriate. On the other hand, if this has already been done, an adaptive sampling technique may be more applicable to find the regions of the design space that have not been covered. These methods often use single and multifidelity regression techniques such as the Kriging and Co-Kriging techniques mentioned earlier, to find the point where the largest variability reduction is expected. For more information, see the work of Park et al.[125], among others.

6.7 Step 6: Iterating as Data is Generated and Requirements Change

Understanding which model is the most appropriate for a particular problem is always an evolving process. The models that are typically used at certain points in the design process and for specific problems became commonly used through repeated evaluation, testing, and accreditation. Many paths can be taken from the previous steps in this framework, some examples of which will be discussed.

6.7.1 Model Representation

If, in Step 4, the calculated R^2 and $RMSE$-values caused models to be scored in a noticeably different way than expected from the initial assessment, then that should be investigated. As discussed previously, there are two primary options for what is occurring: need for troubleshooting, debugging, or other additional verification steps; or a lack of model representation.

There is always the possibility that there is a bug in the code that, when fixed, would cause the model to behave as expected. In such a case, once found, the model set should be re-evaluated to see if it is actually worth considering. This process should be performed first, so that Type I error does not occur, and a valid model is overlooked.

However, in such as case as the model appears to be performing as the model builder would expect, then it is more likely that something about the model lacks the proper representation to tackle the problem at hand. This implies that the initial hypothesis of the model definer, that the model represents the appropriate phenomenology, is incorrect. In certain cases, this could present itself as a difference in the shape of the response surface across the entire design space. However, it could also be a less obvious effect. Type III error could be occurring, such that at least some part of the current design variable range is outside of the range of applicability of the model.

A common example of this is the differences in aerodynamic analysis for different Mach number values. One type of model may be perfectly appropriate in the subsonic region, but once the transonic region is reached for the current configuration, the responses may diverge. This a complicated effect, that depends on the specifics of the current configuration, making it harder to identify.

Another example of where model representation is important related to the type of model in this chapter is the choice of panel stiffener shape. Since the Nastran optimization models in the aircraft wing weight model set represent the panels as unstiffened, a change in the type of stiffener, e.g. they are blade stiffened for this work but could be hat, I, or

Z-stiffened, will show no change in the results. For the HyperSizer models, however, all of these stiffener shapes and more and available to be represented using the smeared stiffness approach. In that case, to get the final optimization result, one of the HyperSizer models must be used.

Experts should be able to account for known regions of applicability, but the transition point may be less obvious or heavily dependent on configuration. In the use case example of this chapter, as shown in Figure 6.8, the aero-static HyperSizer models align with others for aspect ratios below $\approx$10, but then diverge above that point. While the weight estimates do not have such an obvious divergence, there is clearly a division where, for this vehicle design, aeroelasticity is more or less important depending on the specific value of aspect ratio. For one thing, this denotes the importance of choosing appropriate variable ranges. More importantly, it presents a case for how, given the same set of models, the decision-making process could be altered based on the design variable region.

6.7.2 Changing Requirements/Additional Data

The selection of a model or models is dependent on fidelity and efficiency. Multifidelity considerations come into play primarily when the desired fidelity is unattainable due to cost. However, the allowable cost changes based on the problem definition and phase of design. While a minimum allowable fidelity is more of a subjective requirement, it also changes over time accordingly to the change in allowable cost. The increase in allowable cost in more detailed trade studies is, in fact, due to the need for higher fidelity. As requirements change, the Pareto fronts should be revisited to see if the most appropriate model or models has changed.

In addition to changing requirements, the understanding of fidelity through comparative data analysis is based on the currently available amount of data. As such, there can be uncertainty in this assessment when there is little data avilable for all models or a significant discrepancy between the amount of available data between two models, as the correlation

and error metric calculations will be based on some amount of interpolation. As such, it is always desirable to update the assessment any time a significant number of new evaluations is performed.

6.8 Conclusions

Model selection is a process that must often occur prior to a full understanding of the implications. Additionally, since the initial model selection phases are typically heavily-reliant on expert opinion, the possible modeling options and justifications for model selection are often lost in the process. As such, if a reevaluation of the models is required, the process must essentially be restarted from the beginning.

This chapter defines the steps that should be followed, based on the fidelity framework developed in Chapter 4, and the methods developed in Chapter 5, to capture expert opinions, both in the options that are available, but in how they relate to each other in terms of fidelity. Using resolution, abstraction, and scope to define the relative fidelities, the difference between any two models can be described in a way that is more intuitive and clearly defined.

Once a problem is defined **(Step 1)**, applicable modeling options are presented by experts **(Step 2)**, and their relative descriptive fidelities can be explored **(Step 3)**. Based on this, and any available model data **(Step 4)**, the relative fidelities begin to be understood. By comparing fidelity and some representation of model cost, an informed decision can be made based on the minimum fidelity and maximum cost requirements of the current phase of application **(Step 5)**.

If models score differently than experts predicted, troubleshooting or additional research can be performed to find the cause and determine if the model should continue be included for consideration. Otherwise, a model or models can be selected, and new evaluation points can be selected and analyzed. As new data becomes available, requirements change, or new models are brought forward for consideration, the decision-making process

should be revisited **(Step 6)** to reevaluate the most appropriate model selection.

CHAPTER 7

CONTRIBUTIONS, POTENTIAL FOR FUTURE WORK, AND CONCLUSIONS

7.1 Contributions

In this work, a number of issues are tackled related to initial understanding of modeling options to be used for analysis and design. These issues related to understanding model selection, defining fidelity, and how to estimate which model or models is most appropriate, especially in the absence of an abundance of data. This is important since most methods for proving a model's appropriateness, or credibility, require a great deal of data that is not generally available early in the process. Estimating credibility without data, however, adds uncertainty, so the process must be clearly defined and traceable.

Gaining traceable insight into models as early as possible hopes to avoid all types of model errors. If a model's quality can be justified and compared to others based only on a description of its attributes, making the case for the model's accredibility is streamlined, and valid models are more likely to be chosen. Any time spent developing a model that may not be used is risky, but unnecessary development is also avoided in that the process is straightforward enough to iterate any time more model data becomes available, continually reassessing applicability.

7.1.1 Description of Fidelity

Given a problem, experts must attempt to enumerate a list of possible models, and then make an initial selection from that list. Often, the models in that list are not going to be at a point of development that allows for thorough comparison of model data, increasing the reliance on expert opinion. To improve the quality of the expert opinion-based assessment, a literature search was conducted into the field of fidelity definition. Many works have

attempted to define fidelity, and while most of the definitions of model fidelity agree with each other, they are still too general to provide hard levels of insight.

Some of the many fidelity frameworks that have been put forth were enumerated and discussed. Some of these works are useable, but too discipline specific to be generally applied, or high-level, but not as useable. A comparison was made to the field of uncertainty quantification, where the types of uncertainty are defined as aleatory and epistemic, and the sources of uncertainty are always a long case-specific list. Analogously, some, specifically the work of Moon and Hong[83], have identified **resolution** and **abstraction** as fundamentally important to the description of fidelity. However, there is more to the understanding of fidelity than those two atributes.

By compiling the various terms used to describe the aspects of fidelity, filtering out confusing and overlapping terminology, a third group of attributes is defined, represented by the term **scope**. While the resolution, level of detail of the description, and abstraction, amount of simplification to represent the system numerically, are both clearly important and have a complicated interrelationship, scope is often implicitly defined. As design decisions are made, the scope narrows. As resolution increases and abstraction decreases, scope is reduced to make analysis tractable. However, interdisciplinary complications such as aeroelasticity point out that an increased scope would improve the accuracy of the behavior of the model. Scope is interrelated with abstraction in how well that information is passed to the various in-scope components, but if a component is not in the scope of the model, the information cannot be propagated effectively.

Multiple examples are given describing the importance of scope and justifying its inclusion into the description of fidelity. It is shown in the methods developed in Chapter 5 how ignoring an important characteristic of fidelity can lead to a great deal of error in the estimation of fidelity. By comparison of the modeling options for the aircraft use case in Chapter 6, the importance of scope is proven, both in the magnitude of its effect on the response and in the interrelationship with abstraction.

7.1.2 Expert-Elicited Estimation of Model Fidelity

Using resolution, abstraction, and scope as descriptors of the fundamental aspects of fidelity, the focus was turned back to the relative understanding of models purported to aid in an analysis of optimization problem. While those with expertise are gathered to define a model set, any additional insight into the relative order of those models that can be gathered, should be gathered. When expert elicitation is used to understand multifidelity models for model selection uncertainty quantification, the representative metric is the probability that any given model is the highest fidelity in the set. Authorities on the subject are required to explicitly define those probabilities, and the various opinions given are combined. This process obfuscates the problem for two reasons: fidelity is being directly assessed and individuals are spending time dealing with specific values instead of focusing on the aspects of the models.

The first step to clarify this process is using resolution, abstraction, and scope instead of fidelity. Additionally, it is asserted that even those with a great deal of knowledge cannot accurately define the specific ratio between two models with respect to such high-level metrics. As such, the relative resolution, abstraction, and scope of models is defined in terms of whether one model is better, equivalent, or worse than another model.

Given these orderings of models in terms of the three characteristic metrics, normalized scores are automatically calculated, and Kernel Density Estimation is used to convert the scores to distributions describing the combined relative fidelity of each model. By treating these density estimates as random variables, the $P(X > Y)$ can be calculated, and generalized to the entire set to estimate the probability that a given model is the highest fidelity. Additionally, the pairwise assessments can be used to determine the probability that any given model is the 2nd, 3rd, or even lowest fidelity in the set.

Knowing where a model ranks in terms of fidelity improves the understanding of the model set. It allows for a visual manner with which to perform an initial verification that the models fall in the expected order. However, based on the assertion made, the relative

magnitude of the difference between models should be adjusted, but based on quantitative assessment of model data as it becomes available.

7.1.3 Model Fidelity Adjustment Through Comparative Data Analysis

Ideally, the models being put forth for consideration are at some point in the verification and validation process. This means that the validity of the model is not based purely on opinion, and also means that some amount of data may be available, even if manual processes are required. Moving from the descriptive assessment of fidelity, the problem of validation in the absence of experimental data was discussed. It was hypothesized that as the size of the model set increases, agreement between models represents an analogy for validation. As models of varying approaches and fidelities are attempting to estimate the same response, agreement or disagreement helps to strengthen the argument as to the quality of each model. To reiterate the thought experiment, is 100 models are presented, and 99 of them are in high agreement, it is much more likely that the 100th model is a poor representation than that the experts were wrong about the validity of the other 99.

Based on the presumption that, if multiple models are selected, multifidelity Gaussian process regression, or Co-Kriging, may be used to combine them, the work of Toal was discussed[105]. In that work, rules of thumb are put forth for determining when a multifidelity dataset should beget a reliable regression. Specifically, the R^2 and $RMSE$ between any two sets of model data is used to represent the correlation, or agreement of the shape of the data, and error, or residuals, between the magnitudes. Using both of these metrics to create a new set of normalized scores, and using KDE to combine them with the descriptive fidelity scores, an adjusted set of density estimates can be defined.

These adjusted density estimates act to further improve the understanding of the model set similarly to the descriptive assesment. However, by using these correlation and error metrics, additional insight can be gained. When a model set is defined, it is being hypothesized that every given model is representing not just the desired response, but also the

appropriate phenomenology, or behavior, or the problem. Through comparison of the data, models with inadequate representation can be found, even in a high-dimensional problem. The first step to address this should be whether simply troubleshooting is needed, or if the model truly is incapable. However, this allows for iteration and potential down-selection, of the appropriate models based on quantitative justification.

7.1.4 Enabling Multi-Attribute Decision-Making

Model selection is never based purely on fidelity. Going back to the work of Toal[105], the other set of recommendations is used as the basis for appraisal of model efficiency for comparison to the fidelity assessment for informed decision-making. By developing a method for assessing the cost of any model in terms of time, and the relative cost ratios between models, multi-model efficiency can be represented.

Through a method for calculating a multifidelity score and a multi-model efficiency score, the multi-attribute decision-making process can be represented visually using Pareto fronts. The non-dominated multifidelity options are compared to the fidelity and cost of single models. When the desired fidelity is too inefficient, a cost requirement can be overlaid, ruling out single and multi-model combinations that are too expensive, and showing what fidelity can be achieved at the current point in the process. Importantly, as these requirements change, moving from conceptual, to preliminary, to detailed design, a different model or set of models will present as the most favorable option.

Since model cost is an experimental metric, is does not have the same requirements as fidelity. There is no design space to explore in the same way, so even a single runtime can be used to include the cost and efficiency of a given model. This is important since fidelity can be assessed prior to model data generation, so cost should be able to leverage minimal data to enable multi-attribute comparison. Another option discussed but not implemented in this work is to treat the relative dimensionality of various models as an analogous cost. Improving the aspects of fidelity entails increased dimensionality, often by a dramatic amount. If

the change in the number of variables can be assessed, then those numbers can be used to represent the relative cost of each model. This requires additional upfront work from model builders, but could be especially important when all of the possible models have a long development cycle.

7.2 Potential for Future Work

7.2.1 Incorporation of Experimental Data

The methodology developed here estimates model fidelity starting by eliciting orders with respect to resolution, abstraction, and scope. Then, if model data is available, this assessment is adjusted by comparative assessment. One of the reasons given for development of this framework is a common lack of appropriate experimental data. However, even if experimental data exists, on top of likely being sparse, it comes with its own set of assumptions and limitations. This work seeks to take the most advantage of whatever model data is available, so future work could look at the best way to incorporate experimental data when available, one of which is as the highest fidelity data set in a multifidelity regression.

As mentioned earlier in the work, ensemble learning approaches such as bayesian model averaging can be used to generate a set of model probabilities through comparison of model and experimental data. The simplest approach would be to override the other model probabilities with those generated from validation data, however, the methods developed herein provide additional information. Calculating the correlation and error metrics allows for troubleshooting, potential down-selection, and justification of multifidelity combinations. Additionally, finding not just the probability of being the highest fidelity model but the lowest provides additional insight into the model set. On top of that, the pairwise fidelity probabilities are needed for multifidelity scoring.

It is possible that some ensemble learning methods can provide this additional information directly, in which case, the multifidelity scoring approach could be undertaken directly from validation-based probabilities. However, another option would be to simply add the

experimentally-derived scores via kernel density estimation in the same manner as the correlation and error scores. The relative weightings in density estimation could be adjusted since the experimental data should be trusted at least as much as the other scores.

7.2.2 Other Comparative Scoring Methods

Other, more intensive methods for comparative data analysis could be considered in the future. One example would be the re-examination of other goodness-of-fit metrics when appropriate. While R^2 and $RMSE$ were justified for the model sets included in this work, certain data sets could be better understoof by other metrics, or more metrics could just be included to the methodology as a whole.

One additional area of research would pertain directly to the development of multifidelity Gaussian process regressions, or Co-Kriging surrogate models. These models are built by generating a single-model GPR for the lowest fidelity level data, then fitting pairwise Gaussian regressions to the difference between the data sets, with a scaling factor between the models to aid in the fitting process. The specifics of fitting these models is an active area of research.

A typical GPR can generate a covariance matrix, denoting how the value at one point in the design space influences the value at another point in the design space. Multifidelity GPRs can, correspondingly, generate a cross-covariance matrix, which consists of a covariance matrix within each sub-GPR and describes how the various GPRs influence one another in the multifidelity regression.

Future work could aim to see how this information could be used in the comparative data analysis process. The difficulty comes in that, to evalute this, a multifidelity regression must be developed for each permutation of models. However, since this type of multifidelity regression is based on a combination of single-fidelity regression with a number of regressions on pairwise differences between data sets, a method could be developed to efficiently build the regressions from the appropriate set of subregressions.

7.2.3 Efficient Permutation Iteration and Model Compatibilities

It is mentioned several times in this work how the size of the model set can drastically effect the efficiency of the multifidelity scoring methods developed herein. As the size of the set grows linearly, the number of permutations exponentially, drastically increasing the computational requirements. The alleviation method for single attribute scoring involves truncating the size of the combination to be evaluated, which is justified since users are less likely to pick larger model sets for logistical reasons. However, this is a heuristic-based approach that can have miss certain combinations, potentially hiding options from the user that could add to the understanding of the model set.

Additionally, for multi-attribute scoring, only the non-dominated set is retained, lessening the memory requirements. However, this does not reduce the runtime, and no information about combinations near, but not on, the Pareto front is retained. In the future some other factors could be taken into account.

Compatibility checking

While the estimation of fidelity is dependent on a particular, shared, response scenario, there are other values that can be extracted from most models. One example given in the aircraft model set is that of detailed structural data for manufacturing analysis, which can only be generated by the models in set 3 that use HyperSizer. If all ordered model combinations that did not include at least one HyperSizer model were excluded, the number of possible orderings would be drastically reduced.

In addition, the user may wish to exclude combinations between two very similar models. For example, for the wing weight estimation model set, the corresponding Tri and Quad model with the same optimization, aerodynamics, and scope may not represent a desirable combination due to their similarity. If two such models are similar in terms of fidelity and cost, their combination may not present as desirable even if scored. However, if this is known a priori as a combination that should not be evaluated, some amount of time can be

saved by excluding all combinations of those two models.

Tree-Based Approach

The multifidelity scoring metrics are based on multiplying the appropriate sequence of probabilities or efficiency ratios together to achieve the final score. All of these ratios are calculated prior to iteration through permutations, so a more sophisticated approach could be developed to lay out the permutations in a tree. The algorithm would then seek out the highest-valued branches to find the most highly rated permutations instead of simply iterating through all of the possible options.

7.2.4 Adjusted Cost and Efficiency Penalties

The efficiency ratio of

$$E_r = \begin{cases} \frac{1.75}{1+1/C_r} & C_r \in (0, 16/19] \\ \frac{1+C_r}{2} & C_r \in (16/19, \infty) \end{cases}$$

was selected because it is based on the recommended multifidelity comparison method of Toal, and pivots at $C_r = 16/19$, where the upper recommended bound of $f_r = 80\%$ is met. However, further research could investigate adjustments to this cost/reward function. Using a constant of 2 and a pivot point of 1.0 makes the curves line up, however, using a lower number would reward a low C_r more and penalize a high C_r more, as shown in figure 7.1. This allows for the user to change the conservatism with which the ordered model combination efficiency is ranked.

As the process is used and put through its paces with other model sets, specific recommendations other than the ones used here may arise for these settings. For this work, a certain set of values is chosen based on logical assumptions in the process of developing a methodology, and this is presented to make the user aware of which parameters could be changed to effect the conservatism of the efficiency scoring process.

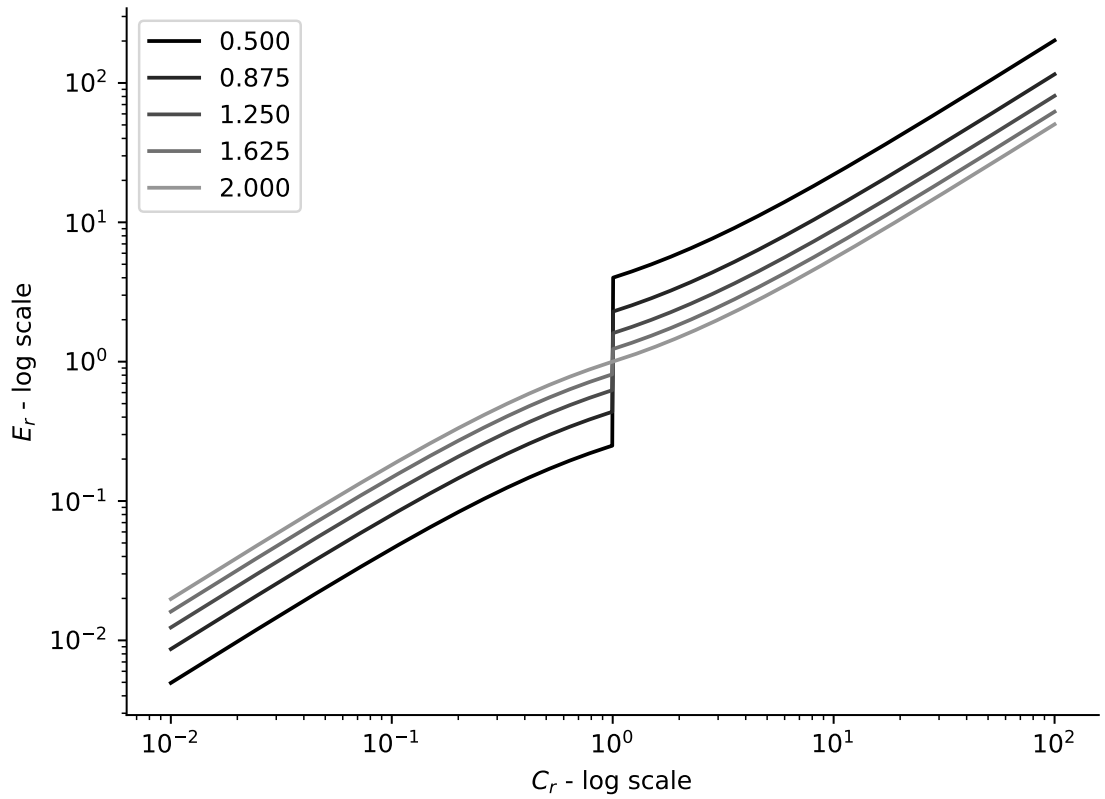


Figure 7.1: E_r with different constants

7.2.5 Regions of Applicability

As discussed previously, when a model is included for consideration, an expert is hypothesizing that not only can the desired response be predicted, but that the appropriate behavior is represented. The methods of the comparative data analysis step can be used to identify when this hypothesis is incorrect across a significant portion of the design space, because an insufficient model should disagree with the other models in the set. Examples of this were the difference between subsonic and transonic for aerodynamic models, or trying to change the panel stiffener type when the shell panels are treated as unstiffened by the optimizer.

When the region of applicability is more subtle, such as the transition between the behavior of low and high aspect ratio wings, or the exact point where the flow field enters the transonic regime, more sophistication is needed to identify the bifurcation point. Given

that the ranges set in problem definition are important to avoiding Type III error, or the application of a valid model to an invalidating application, future research could use a machine learning technique to identify if there ar e different regions of the design space where model agreement varies in comparative data analysis.

If separate regions of applicability are found, separate decision-making problems could take place based on the differing behavior. This could also help to apply bounds within the full variable ranges where a particular model should and shouldn't be used. This would specifically apply to a multifidelity approach, where, in the valid region, sample points from one model should be used to train the regression, but other models are used outside of that region.

7.3 Conclusions

In engineering, understanding comes through gathering data, but prior to the existence of data, the source of data must be selected, which presents a host of potential pitfalls. Models provide all the benefits of gathering data without requirement the time and money to build physical representations of the system, but require abstraction, or simplification, of reality. The initial process of selecting a model is often performed ad hoc, and the decisions based as much on availability as applicability. Even when a thorough exploration of modeling options is performed, after a model is selected, the various options and justifications for selection is often lost.

The framework developed herein provides a way to capture information from experts, not only about which models they think are appropriate, but how they differ in terms general enough that a non-expert can understand and compare. The developed methods provide justification at a point where expert opinion is required, so very little justification is often given, especially justification that is sufficiently captured.

Additionally, this work provides recommendations for how model builders should be thinking about gathering data in order to get their models accredited:

- Data should be gathered throughout development, verification, and validation, even if some part of the process is manual
- As much information as possible should be recorded regarding the cost of generating, analyzing, and post-processing

Fidelity itself is a term that is difficult to define in a usable manner. A great deal of work has been done to attempt to determine how model fidelity should be described and discussed. Resolution and abstraction are terms that have been used to help describe model fidelity, as they are important in defining how well a model represents the system of interest, and have a complex interrelationship that affects model accuracy.

Through literature search, examination of model attributes, and analysis of impact, scope is put forth as the third primary aspect of fidelity. Resolution, abstraction, and scope are the model characteristics that combine to define how well the system is being represented, and are much more linguistically specific and easily understood than fidelity itself. As such, these terms should be used to aid in the discussion of model fidelity moving forward.

The process of estimating the relative fidelity of models in a set is crucial to selecting a model or models from that set. Through the methods developed in Chapter 5, fidelity can be better understood through model probabilities and a multifidelity order scoring process. Additionaly, the understanding of model cost is important not only for the generation of a multifidelity surrogate model, but for the selection of any model.

As model data becomes available, it should be used to update the relative distances betwen models, identify insufficiencies, and aid in verification and validation. The understanding of fidelity, whether with model data or without, should then be joined with a representation of the cost of evaluating each model to find the non-dominated set of multifidelity options and compare them to the fidelity and efficiency of the individual models.

Given an infinite computational budget, the highest fidelity model would be most desirable and multifidelity options would only be considered to increase robustness or reduce

uncertainty. However, the cost per evaluation is often a limiting factor, so, at various points in the design process, a different model or ordered combination of models becomes the most desirable option for continued development and evaluation.

Based on a selection, new design points are selected and analyzed by the model or models based on their relative costs and which points have already been evaluated. The process of understanding the fidelity and efficiency of multifidelity modeling options, decsribed in this framework, should not be performed just once, but revisited as more data becomes available, new models are presented, or requirements change.

REFERENCES

[1] G. E. P. Box, *Robustness in the strategy of scientific model building*, R. L. Launer and G. N. Wilkinson, Eds. New York, New York, USA: Academic Press, Inc., 1979, pp. 201–236.

[2] D. P. Thunnissen, "Propagating and Mitigating Uncertainty in the Design of Complex Multidisciplinary Systems," PhD thesis, California Institute of Technology, 2005.

[3] J. Mullins and H. Kim, "Fidelity Forward Multidisciplinary Analysis and Optimization : Harnessing the Power of High Fidelity CAD and CAE Tools in Conceptual Design," *Director*, no. September, pp. 1–20, 2008.

[4] J. Ceisel, P. Witte, and T. Carr, "A Non-Weight Based, Manufacturing Influenced Design (MInD) Methodology for Preliminary Design," *28th International Congress of the Aeronautical Sciences*, pp. 1–10, 2012.

[5] T. R. Milner, "A Risk-Informed Manufacturing Influenced Design Framework for Affordable Launch Vehicles," PhD thesis, Georgia Institute of Technology, 2016.

[6] C. H. Lee, "Bayesian Collaborative Sampling for Multidisciplinary Design," Ph.D. Georgia Institute of Technology, 2012, p. 233, ISBN: 9781600869303.

[7] N. Courrier, P. A. Boucard, and B. Soulier, "The use of partially converged simulations in building surrogate models," *Advances in Engineering Software*, vol. 67, pp. 186–197, 2014.

[8] L. Le Gratiet, "Bayesian Analysis of Hierarchical Multifidelity Codes," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 1, no. 1, pp. 244–269, 2013.

[9] M. C. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, vol. 87, no. 1, pp. 1–13, 2000.

[10] D. L. Allaire, K. E. Willcox, and O. Toupet, "A Bayesian-Based Approach to Multifidelity Multidisciplinary Design Optimization," pp. 2010–9183, 2010.

[11] L. W. T. Ng and K. E. Willcox, "Multifidelity approaches for optimization under uncertainty," *International Journal for Numerical Methods in Engineering*, vol. 100, no. 10, pp. 746–772, 2014.

[12] N. M. Alexandrov, E. Nielsen, R. Lewis, and W. Anderson, "First-order model management with variable-fidelity physics applied to multi-element airfoil optimization," in *8th Symposium on Multidisciplinary Analysis and Optimization*, 2000.

[13] N. M. Alexandrov, R. M. Lewis, C. R. Gumbert, L. L. Green, and P. A. Newman, "Approximation and Model Management in Aerodynamic Optimization with Variable-Fidelity Models," *Journal of Aircraft*, vol. 38, no. 6, pp. 1093–1101, 2001.

[14] A. I. March, "Multifidelity Methods for Multidisciplinary System Design," Doctor of Philosophy, Massachusetts Institute of Technology, 2012, p. 220.

[15] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[16] J. A. Sokolowski and C. M. Banks, *Principles of Modeling and Simulation: A Multidisciplinary Approach*. John Wiley & Sons, 2011, ISBN: 9780470403563.

[17] G. Murphy, *Similitude in engineering*. Ronald Press Co., 1950.

[18] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*. McGraw-Hill New York, 1991, vol. 2, ISBN: 0070592926.

[19] H. Chestnut, *Systems Engineering Tools*. New York: John Wiley and Sons, Inc., 1965.

[20] G. E. Dieter and L. C. Schmidt, *Engineering Design*. McGraw-Hill New York, 2000, vol. 3.

[21] W Humphries Sr, J. C. Blair, R. S. Ryan, and L. A. Schutzenhofer, "Launch Vehicle Design Process: Characterization, Technical Integration, and Lessons Learned," *Technical Memorandum*, no. May, pp. 1–261, 2001.

[22] W. Heisenberg, "Physics and Philosophy The Revolution in Modern Science," *Book*, 1958.

[23] J. G. Adair, "The Hawthorne Effect: A Reconsideration of the Methodological Artifact," *Journal of Applied Psychology*, vol. 69, no. 2, pp. 334–345, 1984.

[24] R. de la Fuente-Fernández, T. J. Ruth, V. Sossi, M. Schulzer, D. B. Calne, and A. J. Stoessp, "Expectation and Dopamine Release : Mechanism of the Placebo Effect in Parkinson ' s Disease," vol. 293, no. August, pp. 1164–1167, 2001.

[25] T. Mytkowicz, P. Sweeney, M. Hauswirth, and A. Diwan, "Observer Effect and Measurement Bias in Performance Analysis," vol. 972, no. June, 2008.

[26] J. D. Anderson Jr, *Fundamentals of Aerodynamics*, 3rd ed. New York: McGraw-Hill Higher Education, 2001, vol. Third Edit, p. 982, ISBN: 0073398101.

[27] I. E. Alber, *Aerospace Engineering on the Back of an Envelope*. 2015, vol. 1, p. 326, ISBN: 9788578110796.

[28] T. G. Trucano, L. P. Swiler, T. Igusa, W. L. Oberkampf, and M. Pilch, "Calibration, validation, and sensitivity analysis: What's what," *Reliability Engineering and System Safety*, vol. 91, no. 10-11, pp. 1331–1357, 2006.

[29] P. J. Roache, "Verification of Codes and Calculations," *AIAA JOURNAL*, vol. 36, no. 5, 1998.

[30] Department of Defense, *MIL-STD-3022 Documentation of Verification, Validation, and Accreditation (VV&A) for Models and Simulations*, 2008.

[31] S. Lacaze and S. Missoum, "Bayesian calibration using fidelity maps," in *Proceedings of the 11th International Conference on Structural Safety & Reliability*, 2013, ISBN: 9781138000865.

[32] M. C. Kennedy and A. O'Hagan, "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001.

[33] S. Sankararaman and S. Mahadevan, "Integration of Model Verification, Validation, and Calibration for Uncertainty Quantification in Engineering Systems," *Reliability Engineering & System Safety*, vol. 138, pp. 194–209, 2015.

[34] Y. Heo, D. J. Graziano, L. Guzowski, and R. T. Muehleisen, "Evaluation of calibration efficacy under different levels of uncertainty," *Journal of Building Performance Simulation*, vol. 8, no. 3, pp. 135–144, 2015.

[35] Department of Defense, *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide*, 2011.

[36] D. Bigoni, "Uncertainty Quantification with Applications to Engineering Problems," Ph.D. Technical University of Denmark, 2014.

[37] W. L. Oberkampf and C. J. Roy, *Verification and Validation in Scientific Computing*. Cambridge University Press, 2010, p. 791, ISBN: 9780521113601/ 9780511906725.

[38] W. L. Oberkampf, J. C. Helton, C. A. Joslyn, S. F. Wojtkiewicz, and S. Ferson, "Challenge problems: Uncertainty in system response given uncertain parameters," in *Reliability Engineering and System Safety*, vol. 85, 2004, pp. 11–19, ISBN: 1505844452.

[39] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Structural Safety*, vol. 31, pp. 105–112, 2008.

[40] B. E. Robertson, "A Hybrid Probabilistic Method To Estimate Design Margin," Ph.D. Georgia Institute of Technology, 2013, pp. 1–344.

[41] R. E. Melchers, *Structural reliability analysis and prediction*, 2nd. Joh, 1999.

[42] J. Riera, M. Rocha, and R. Ramos de Menezes, "On the consideration of phenomenological uncertainty," *Nuclear Engineering and Design*, vol. 175, no. 3, pp. 259–265, 1997.

[43] J. Ludewig, "Models in software engineering – an introduction," *Software and Systems Modeling*, vol. 2, pp. 5–14, 2003.

[44] NASA (National Aeronautics and Space Administration), "NASA strategic plan," Tech. Rep., 2014.

[45] N. Bennett and G. J. Lemoine, *What VUCA really means for you*, 2014.

[46] S. Bernstein, "Methodology for Interoperability-Enabled Adaptable Strategic Fleet Mix Planning," PhD thesis, Georgia Institute of Technology, 2017.

[47] D. C. Montgomery, *Design and analysis of experiments*, 7th ed. New York, New York, USA: John Wiley & Sons, Inc., 2009.

[48] J. Antony, *Design of experiments for engineers and scientists*, 1st ed. Elsevier, 2003.

[49] R. H. Myers and D. C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed. New York, New York, USA: Wiley, 2002.

[50] J. P. C. Kleijnen, *Kriging metamodeling in simulation: A review*, 2009.

[51] R. B. Gramacy and H. K. H. Lee, "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1119–1130, 2008.

[52] L. Le Gratiet and J. Garnier, "Recursive Co-Kriging Model for Design of Computer Experiments With Multiple Levels of Fidelity," *International Journal for Uncertainty Quantification*, vol. 4, no. 5, pp. 365–386, 2014.

[53] J. D. Martin and T. W. Simpson, "Use of Kriging Models to Approximate Deterministic Computer Models," *AIAA Journal*, vol. 43, no. 4, pp. 853–863, 2005.

[54] M. E. Riley and R. V. Grandhi, "Quantification of Modeling-Induced Uncertainties in Simulation-Based Analyses," *AIAA Journal*, vol. 52, no. 1, pp. 195–202, 2014.

[55] M. E. Riley and R. V. Grandhi, "Quantification of model-form and predictive uncertainty for multi-physics simulation," in *Computers and Structures*, vol. 89, Pergamon, 2011, pp. 1206–1213, ISBN: 0045-7949.

[56] W. L. Oberkampf, T. G. Trucano, and C. Hirsch, "Verification, validation, and predictive capability in computational engineering and physics," *Applied Mechanics Reviews*, vol. 57, no. 5, p. 345, 2004.

[57] C. G. Soares, "Quantification of model uncertainty in structural reliability," in *Probabilistic Methods for Structural Design*, C. G. Soares, Ed., 56th ed., Dordrecht: Springer, 1997, ch. 2, pp. 17–37, ISBN: 0047701003770.

[58] G. Lin, D. Engel, and P. Eslinger, "Survey and evaluate Uncertainty Quantification Methodologies," *U.S. Department of Energy*, no. February, pp. 1–26, 2012.

[59] E. Zio and G. E. Apostolakis, "Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories," *Reliability Engineering and System Safety*, vol. 54, no. 2-3, pp. 225–241, 1996.

[60] E. Borgonovo and E. Plischke, *Sensitivity analysis: A review of recent advances*, 2016.

[61] H. Rabitz, "Systems Analysis at the Molecular Scale," *American Institute for the Advancement of Science1*, vol. 246, no. 4927, pp. 221–226, 1989.

[62] F. Ferretti, A. Saltelli, and S. Tarantola, "Trends in sensitivity analysis practice in the last decade," *Science of the Total Environment*, vol. 568, pp. 666–670, 2016.

[63] A. Saltelli, K. Chan, and E. M. Scott, *Sensitivity Analysis - Numbers for policy: Practical problems in quantification*, 1st ed. New York, New York, USA: Wiley, 2000.

[64] B. Liang and S. Mahadevan, "Error and Uncertainty Quantification and Sensitivity Analysis in Mechanics Computational Models," *International Journal for Uncertainty Quantification*, vol. 1, no. 2, pp. 147–161, 2011.

[65] A. Saltelli, "Sensitivity analysis for importance assessment," *Risk Analysis*, vol. 22, no. 3, pp. 579–590, 2002.

[66] B. Iooss and M. Ribatet, "Global sensitivity analysis of computer models with functional inputs," *Reliability Engineering & System Safety*, vol. 94, no. 7, pp. 1194–1204, 2009.

[67] L. Uusitalo, A. Lehikoinen, I. Helle, and K. Myrberg, *An overview of methods to evaluate uncertainty of deterministic models in decision support*, 2015.

[68] N. E. Lane and E. A. Alluisi, "Fidelity and validity in distrubuted interactive simulation : questions and answers," *Security*, 1992.

[69] D. F. Freeman, "A Product Family Design Methodology Employing Pattern Recognition," PhD thesis, Georgia Instititue of Technology, 2013.

[70] M. P. Bailey and W. G. Kemple, "The scientific method of choosing model fidelity," in *Proceedings of the 1992 Winter Simulation Conference*, New York, New York, USA: ACM Press, 1992, pp. 791–797, ISBN: 0780307984.

[71] Z. C. Roza, D. C. Gross, and S. Y. Harmon, "M&S VV&A RPG Special Topic: Fidelity," in *Spring simulation Interoperability Workshop*, 2000, pp. 1–10.

[72] D. C. Gross, "Report from the Fidelity Implementation Study Group," p. 88, 1999.

[73] I. Park, H. K. Amarchinta, and R. V. Grandhi, "A Bayesian approach for quantification of model uncertainty," *Reliability Engineering & System Safety*, vol. 95, no. 7, pp. 777–785, 2010.

[74] M. Price, S. Raghunathan, and R. Curran, *An integrated systems engineering approach to aircraft design*, 2006.

[75] J. Melorose, R. Perroy, and S. Careas, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proceedings of the IEEE*, vol. 1, no. 1, pp. 148–175, 2015.

[76] C. O'Sullivan, J. Dingliana, T. Giang, and M. K. Kaiser, "Evaluating the visual fidelity of physically based animations," *ACM Transactions on Graphics*, vol. 22, no. 3, p. 527, 2003.

[77] M. E. Riley and R. V. Grandhi, "Quantification of Modeling-Induced Uncertainties in Simulation-Based Design," in *AIAA/ASME/ASCE/AHSI/ASC Structural Dynamics and Materials Conference*, Honolulu, HI, 2012, pp. 1–18.

[78] R. Rebba, "Model validation and design under uncertainty," Ph.D. Vanderbilt University, 2005.

[79] K. K. Gupta and J. L. Meek, *Finite Element Multidisciplinary Analysis*, 2nd ed. Reston, Virginia: American Institute of Aeronautics and Astronautics, 2003.

[80] J. Roskam, "Part V: Component Weight Estimation," in *Airplane Design*, 4th ed., Lawrence, Kansas: Design, Analysis, and Research Corporation, 2017, ch. 5.

[81] M. Roza, J. Voogd, and P. Van Gool, "FIDELITY CONSIDERATIONS FOR CIVIL AVIATION DISTRIBUTED SIMULATIONS," *AIAA Modeling and Simulation Technologies Conference*, 2000.

[82] Z. C. Roza, D. C. Gross, and S. Y. Harmon, "Report out of the Fidelity Experimentation Implementation Study Group," in *SISO Spring 2000 Simulation Interoperability Workshop*, 2000.

[83] I. C. Moon and J. H. Hong, "Theoretic interplay between abstraction, resolution, and fidelity in model information," *Proceedings of the 2013 Winter Simulation Conference - Simulation: Making Decisions in a Complex World, WSC 2013*, no. Roza 2004, pp. 1283–1291, 2013.

[84] Z. C. Roza, "Simulation fidelity theory and practice," PhD thesis, 2005, ISBN: 9040725691.

[85] G. E. P. Box and R. D. Meyer, "An Analysis of Unreplicated Fractional Factorials," *Technometrics*, vol. 28, no. 1, pp. 11–18, 1986.

[86] T. J. Rudman and K. L. Austad, "The Centaur Upper Stage Vehicle," no. December, 2002.

[87] G. P. Sutton and O. Biblarz, *Rocket Propulsion Elements*, 8th. Hoboken, New Jersey: John Wiley & Sons, Inc., 2010, p. 768, ISBN: 978-0-470-08024-5.

[88] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[89] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[90] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[91] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cour-

napeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research2*, vol. 12, pp. 2825–2830, 2011.

[92] J. T. Vanderplas, *Python data science handbook : essential tools for working with data*, First edit. 2016, p. 529, ISBN: 9781491912058.

[93] E. Jones, T. Oliphant, P. Peterson, and Others, *SciPy: Open source scientific tools for Python*, 2001.

[94] P. Safarian, "Finite Element Modeling and Analysis Validation," in *Femap Symposium Series*, 2015.

[95] W. C. Young, R. G. Budynas, and A. M. Sadegh, *Roark's Formulas for Stress and Strain*, 8th. 2011, ISBN: 0071742476.

[96] ASTM, *ASTM Structural Steel Wide Channel I- Beam Cross Section Properties*.

[97] MSC Software Corporation, *MSC Nastran Multidisciplinary Structural Analysis Software*, 2017.

[98] *MSC Nastran 2017.1 Quick Reference Guide*. MacNeal-Schwendler Corporation, 2017.

[99] *MSC Nastran 2017.1 Linear Static Analysis User's Guide*. MacNeal-Schwendler Corporation, 2017.

[100] MSC Software Product Marketing, *Choosing the Right Finite Element — MSC Nastran*, 2013.

[101] The HDF Group, *What is HDF5?*

[102] MSC Software Product Marketing, *HDF5: A Useful Enhancement for MSC Nastran and Patran*, 2018.

[103] A. Collette, *Python and HDF5: Unlocking Scientific Data*, 1st ed. O'Reilly Media, 2013, p. 152, ISBN: 1491944994.

[104] R. Sjoegren, *pyDOE2 1.1.2: Design of experiments for Python*.

[105] D. J. J. Toal, "Some considerations regarding the use of multi-fidelity Kriging in the construction of surrogate models," *Structural and Multidisciplinary Optimization*, vol. 51, no. 6, pp. 1223–1245, 2015.

[106] D. Raymer, *Aircraft Design: A Conceptual Approach, Fifth Edition*. Washington, DC: American Institute of Aeronautics and Astronautics, Inc., 2012, ISBN: 978-1-60086-911-2.

[107] R. G. Budynas and K. J. Nisbett, *Shigley's Mechanical Engineering Design*, 10th. McGraw-Hill Education, 2014, p. 1104, ISBN: 9780073398204.

[108] A. I. J. Forrester, A. Sobester, and A. J. Keane, *Engineering Design Via Surrogate Modelling : A Practical Guide*. J. Wiley, 2008, p. 210, ISBN: 9780470770801.

[109] F. Biscani, D. Izzo, W. Jakob, M. Märtens, A. Mereta, C. Kaldemeyer, S. Lyskov, S. Corlay, B. Pritchard, K. Manani, J. Mabille, T. Miąsko, A. Huebl, Jakirkham, Hulucc, Polygon, Z. Fu, T. G. Badger, M. Nimier-David, L. Č. Zajc, J. Adler, J. Travers, J. Lee, J. Jordan, I. Smirnov, H. Nguyen, F. Lema, E. O'Leary, and A. Mambrini, "Esa/pagmo2: Pagmo 2.10," 2019.

[110] M. Jensen, "Reducing the Run-Time Complexity of Multiobjective EAs: The NSGA-II and Other Algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 5, pp. 503–515, 2003.

[111] D. P. Schrage, "Technology for Rotorcraft Affordability Through Integrated Product/Process Development (IPPD)*," Tech. Rep., 1999.

[112] National Aeronautics and Space Administration, *NASA Common Research Model*.

[113] G. Kenway, G. Kennedy, and J. Martins, "Aerostructural optimization of the Common Research Model configuration," in *15th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Reston, Virginia: American Institute of Aeronautics and Astronautics, 2014, ISBN: 978-1-62410-283-7.

[114] J. Gloudemans, R. McDonald, M. Moore, A. Hahn, B. Fredericks, and A. Gary, *OpenVSP*.

[115] A. A. S. M. Inc., *ASM Material Data Sheet: Aluminum 6061-T6*.

[116] J. Bolognese, *FEMCI Book - Unit Consistency In Nastran*, 2008.

[117] J. A. Corman, N. Weston, C. Friedland, D. N. Mavris, and T. W. Laughlin, "Rapid Airframe Design Environment (RADE): A Parametric Multi-Fidelity Approach to Conceptual Airframe Design," in *2018 AIAA Modeling and Simulation Technologies Conference*, Kissimmee, Florida, 2018, ISBN: 978-1-62410-528-9.

[118] A. Sudol, "A Methodology for Modeling the Verification, Validation, and Testing Process for Launch Vehicles," PhD thesis, Georgia Institute of Technology, 2015.

[119] C. S. Collier, *HyperSizer Composite Analysis and Structural Sizing Software*, 2016.

[120] Collier Research Corporation, *Smeared Stiffness Image*.

[121] MSC Software Corporation, *MSC Nastran 2017.1 Design Sensitivity and Optimization User's Guide*. MacNeal-Schwendler Corporation, 2017.

[122] M. Drela and H. Youngren, *AVL: Athena Vortex Lattice*.

[123] MSC Software Corporation, *MSC Nastran 2017.1 Aeroelastic Analysis User's Guide*. MacNeal-Schwendler Corporation, 2017.

[124] C. B. Barber and D. P. Dobkin, "The Quickhull Algorithm for Convex Hulls," Tech. Rep., 1996.

[125] C. Park, R. T. Haftka, and N. H. Kim, "Remarks on multi-fidelity surrogates," *Structural and Multidisciplinary Optimization*, vol. 55, no. 3, pp. 1029–1050, 2017.